Tampereen yliopisto

Aleksi Jokela

DATAN LAADUN MITTAAMINEN JA ARVIOINTI

Tekniikan ja luonnontieteiden tiedekunta
Diplomityö
Syyskuu 2019

TIIVISTELMÄ

ALEKSI JOKELA: Datan laadun mittaaminen ja arviointi
Tampereen yliopisto
Diplomityö, 104 sivua, 5 liitesivua
Syyskuu 2019
Tietojohtamisen diplomi-insinöörin tutkinto-ohjelma
Pääaineet: Tiedon ja osaamisen hallinta sekä informaatioanalytiikka
Tarkastajat: Professori Samuli Pekkola ja yliopisto-opettaja Ilona Ilvonen

Avainsanat: datan laatu, datan laadun mittaaminen, datan laadun arviointi

Dataresurssien kasvavan määrän ja monimutkaisuuden vuoksi datan laadunhallinnasta on muodostunut tärkeä menestystekijä yrityksille. Laadukkaan datan tärkeys yrityksen päätöksenteossa kasvaa, mutta samalla kasvaa myös haaste datan laadukkuuden varmistamiseksi. Dataa virtaa jatkuvasti yritykseen eri lähteistä, järjestelmistä ja käyttäjiltä, minkä myötä datan määrä kasvaa päivittäin. Datan laadun mittaamisen ymmärtämiseksi on huomioitava neljä asiaa. Miten data, laatu ja mittaaminen ymmärretään sekä miten nämä kolme ensimmäistä liittyvät toisiinsa. Näiden lisäksi tässä työssä esitetään myös arvioinnin merkitys datan laadun diagnosoinnissa.

Diplomityö toteutettiin tapaustutkimuksena suomalaiselle ICT-alan yritykselle. Tutkimuksen tarkoituksena oli selvittää, miten kohdeyrityksen datan laatua voidaan mitata ja arvioida. Tutkimuksen empiriassa käytettiin yhdistelmämenetelmää, joka viittaa määrällisten ja laadullisten tiedonkeruu- ja analysointimenetelmien hyödyntämiseen. Puolistrukturoitujen haastatteluiden avulla kerättyä laadullista ja määrällistä aineistoa triangluoitiin kohdeyrityksen tietokannasta saatavalla määrällisellä Master asiakasdatalla. Aineistojen analysoinnissa hyödynnettiin luokittelua, toistuvuuden laskemista, objektiivisia mittareita, roolien etäisyyksien analysointia ja vertailuanalyysiä. Empiriassa käytettiin Hybridi-arviointimenetelmää, johon sisällytettiin arviointitoimintoja kohdeyrityksen laatuongelmien ja tavoitteiden mukaisesti.

Tutkimuksen tuloksista saatiin numeerisen laadun tason lisäksi myös laadullisia kehitystoimenpiteitä ja haasteita. Numeeriset tulokset antoivat yleiskuvan laadun tasosta eri ulottuvuuksien avulla, kun taas laadullisten tulosten myötä pystyttiin tunnistamaan konkreettisia ongelmakohtia. Ongelmakohtia havaittiin mm. alkuvaiheen rekisteröintiprosesseissa ja avainarvojen hyödyntämisen laajuudessa. Subjektiivisen ja objektiivisen mittaamisen vertailuanalyysin tulokset olivat puolestaan yllättävän lähellä toisiaan. Tutkimuksen toteuttamisen myötä nähtiin, että Hybridi-arviointimenetelmä on potentiaalinen menetelmä datan laadun arviointiin etenkin siihen sisällytettävien arviointitoimintojen joustavuuden myötä.

ABSTRACT

ALEKSI JOKELA: Data quality measurement and assessment
Tampere University
Master of Science Thesis, 104 pages, 5 Appendix pages
September 2019
Master's Degree Programme in Information and Knowledge Management
Majors: Knowledge Management and Information Analytics
Examiners: Professor Samuli Pekkola and University Instructor Ilona Ilvonen

Keywords: data quality, data quality measurement, data quality assessment

Due to the increasing amount and complexity of data resources, data quality management has become an important success factor for companies. The importance of high-quality data in business decision-making is growing, but at the same time, the challenge of ensuring data quality is increasing. Data is constantly flowing to the company from various sources, systems and users, increasing the amount of data every day. To understand data quality measurement, there are four things to consider. How data, quality, and measurement are understood and how the first three relate to each other. In addition to these, the importance of assessment in diagnosing data quality is also presented in this work.

Master's thesis was conducted as a case study for a Finnish ICT-company. The objective of this research was to find out how the target company's data quality can be measured and assessed. A combination method was used in empirical research, which refers to the utilization of quantitative and qualitative data collection and analysis methods. Qualitative and quantitative data that was collected through semi-structured interviews were triangulated with quantitative Master customer data from the target company's database. The empirical material was analyzed using classification, recurrence calculation, objective metrics, role gap analysis, and comparison analysis. The Hybrid assessment method was utilized in empirical research, which included assessment activities related to the target company's quality problems and objectives.

Thesis results provided numerical data quality level and qualitative development procedures and challenges. The numerical results provided an overview of the data quality level through various quality dimensions, whereas the qualitative results identified specific quality problems. Examples of issues identified were in early-stage registration processes and the extent of key-value exploitation. The comparison analysis results of the subjective and objective measurements were surprisingly close to each other. It was also seen from the thesis that the Hybrid assessment method is a potential method for data quality assessment, particularly due to the flexibility of the assessment functions which can be included in it.

ALKUSANAT

”Viimeistelen diplomityön vaihto-opiskelun aikana Prahassa.” Tämä ei aivan toteutunut, enkä omakohtaisten kokemusten myötä voi suositella sen yrittämistä. Nyt muutama kuukausi vaihto-opiskelun jälkeen voin kuitenkin todeta, että tämä projekti on saatu päätökseen.

Kiitos etenkin diplomityön kohdeyritykselle, jolta sain kiinnostavan, motivoivan ja omiin työtehtäviin kohdistuvan aiheen tutkittavaksi. Jos aihe ei olisi ollut kiinnostava, niin en todennäköisesti olisi vielä kirjoittamassa alkusanoja. Haluan samalla kiittää kohdeyritystä myös ensimmäisen oman alan työpaikan tarjoamisesta, mikä toimi hyvänä pohjana uran jatkamiselle.

Kiitos kuuluu myös kohdeyrityksen kahdelle ohjaajalle, joilta sain neuvoja yrityksen tavoitteiden ja tutkimuksen yhteensovittamisessa. Toivottavasti diplomityö tarjoaa vaadittua ymmärrystä datan laadun mittaamisen ja arvioinnin keskeisistä menetelmistä. Toinen ohjaajista toimi lisäksi työparina ja mentorina, mistä olen erittäin kiitollinen.

Yliopiston ohjaajalle kiitos etenkin tieteellisen tutkimuksen toteuttamisen auttamisessa. Moni asia oli aluksi epäselvää, mutta sain apua aina tarvittaessa. Kiitos kuuluu myös opiskelutovereilleni, joilta sain lisämotivaatiota työn viimeistelemiseen. Lopuksi haluan vielä kiittää perhettäni jatkuvasta tuesta.

Tampereella, 28.9.2019

Aleksi Jokela

SISÄLLYSLUETTELO

KUVALUETTELO

TAULUKKOLUETTELO

1. JOHDANTO

Suurin osa päätöksistä perustuu dataan, minkä vuoksi huonoon dataan perustuva päätös voi johtaa negatiivisiin vaikutuksiin. Datan laadun arviointi voi auttaa päättäjiä tuntemaan datan nykytilan ja siten niiden tekemän päätöksen laadun. (Aljumaili et al. 2016) Informaatioteknologian kehitys on auttanut organisaatiota keräämään ja varastoimaan dataa enemmän kuin koskaan aiemmin (Watts et al. 2009). Dataresurssien kasvavan määrän ja monimutkaisuuden vuoksi datan laadunhallinnasta on muodostunut tärkeä menestystekijä yrityksille. Korkealaatuinen data tukee sujuvia toimintoja, mahdollistaa dataohjautuvan päätöksenteon ja edistää kilpailuedun saavuttamista. Vastavuoroisesti heikkolaatuinen data aiheuttaa organisatorista tehottomuutta ja pääoman menetyksiä. (Shankaranarayanan & Even 2007; Heinrich et al. 2018a)

Yritykset tarvitsevat korkealaatuista dataa varastoista, toimittajista, asiakkaista, myyjistä sekä muista tärkeistä yritystiedoista, jotta ne voivat tehokkaasti hyödyntää niiden analysointiohjelmistoja tuottaakseen tarkkoja tuloksia. Mikä tahansa datan laatuongelma voi johtaa virheellisiin analyyseihin, mikä puolestaan voi aiheuttaa vakavia seurauksia. Validin datan tärkeys yrityksen päätöksenteossa kasvaa, mutta samalla kasvaa myös haaste datan validiteetin varmistamiseksi. Dataa virtaa jatkuvasti yritykseen eri lähteistä, järjestelmistä ja käyttäjiltä, minkä myötä datan määrä kasvaa päivittäin. (Andreescu et al. 2014) Datan määrän kasvaessa kasvaa myös sen hallinnan monimutkaisuus ja huonon datan laadun riskit (Watts et al. 2009).

Datan laadunhallinnan ja päätöksenteon tukemiseksi on keskeistä arvioida datan laadun tasoa mittareiden avulla. Jos mittareita ei ole kuitenkaan määritelty riittävän hyvin, niin ne saattavat johtaa vääriin päätöksiin ja taloudellisiin menetyksiin. (Heinrich et al. 2018a) Samalla kun tutkimukset ja käytännöt ovat huomanneet hyvin perusteltujen datan laatumittarien merkityksen, niin monista datan laatumittareista puuttuu kuitenkin asianmukaiset metodiset perustat. Useat mittarit kehitetään ad hoc -periaatteella tiettyjen ongelmien ratkaisemiseksi tai ne ovat hyvin subjektiivisia. (Pipino et al. 2002; Heinrich et al. 2009; Heinrich et al. 2018a)

Yritysten on käsiteltävä molempia sekä datan parissa työskentelevien henkilöiden subjektiivisia näkemyksiä, että kyseiseen dataan liittyviä objektiivisia mittauksia (Pipino et al. 2002). Mittaaminen ja arviointi eroavat toisistaan, sillä mittauksista saadaan arvoja, joita arvioinnissa tarkastellaan tarvittavien datan laadun kehittämistoimenpiteiden määrittämiseksi (Woodall et al. 2013). Datan laadun mittaamisen ymmärtämiseksi on huomioitava neljä asiaa. Miten data, laatu ja mittaaminen ymmärretään sekä miten nämä kolme

ensimmäistä liittyvät toisiinsa. (Sebastian-Coleman 2013 s. 2) Näiden lisäksi tässä työssä esitetään myös arvioinnin merkitys datan laadun diagnosoinnissa.

1.1 Tutkimuksen merkityksen perustelu

Kasvavan datamäärän myötä on alettu kiinnittämään yhä enemmän huomioita datan laatuun. Yrityksellä voi olla paljon dataa, mutta sen hyödynnettävyys saattaa olla alhaista huonon datan laadun vuoksi. Kohdeyrityksen kannalta tutkimuksen aihe on ajankohtainen, sillä yrityksessä on panostettu viime aikoina entistä enemmän datan hallintaan. Datan laatua ja sen hallintaa voidaan pitää yhtenä datan hallinnan osa-alueena, minkä myötä aihe linkittyy laajempaan strategiseen tavoitteeseen. Kohdeyrityksellä ei ollut ennen työn aloittamista yhtenäisiä menetelmiä datan laadun mittaamiseen ja arviointiin, sillä datan laatua saatettiin diagnosoida eri tavoin eri työntekijöiden toimesta.

Tutkimuksen teorian tarkoituksena on tarjota ymmärrystä datan laadun mittaamisen ja arvioinnin keskeisimmistä perusperiaatteista ja menetelmistä. Tutkimuksen tavoitteena ei ole luoda uutta datan laadun mittaamisen ja arvioinnin teoriaa, vaan tuoda esiin alan parhaimmat käytännöt ja soveltaa niitä käytännössä. Empiriassa toteutetaan menetelmien käytännön testaaminen, ja työn lopussa arvioidaan niiden sopivuutta kohdeyrityksen tapauksessa.

Tutkimuksen yhtenä päätavoitteista voidaankin pitää mahdollisimman suurta käytännöllistä kontribuutiota ja liiketoiminnallista hyötyä kohdeyritykselle. Datan laadun mittaamisen ja arvioinnin menetelmien soveltamisen myötä kohdeyritys voi kehittää omia toimintoja datan laatuongelmien mittaamiseen, arviointiin ja seurantaan. Henkilökohtaisesta näkökulmasta työn tekeminen antoi paljon lisäymmärrystä kiinnostavasta aiheesta, joka on merkityksellinen myös työtehtävien näkökulmasta. Tutkimuksesta voivat hyötyä myös muut yritykset, joiden tavoitteena on kasvattaa ymmärrystä datan laadun mittaamisen ja arvioinnin menetelmistä.

1.2 Tutkimusongelma ja tutkimuskysymykset

Diplomityön tarkoituksena on selvittää, miten kohdeyrityksen datan laatua voidaan mitata ja arvioida. Tämän selvittämiseksi työssä esitetään datan laadun mittaamisen ja arvioinnin teoriaa, jonka menetelmiä testataan käytännössä. Teoriassa keskitytään yhdistämään data, laatu, mittaaminen ja arviointi yhtenäiseksi kokonaisuudeksi sekä esittämään mittaamisen ja arvioinnin keskeisimmät menetelmät.

Työn painopiste on datan laadun mittaamisessa ja arvioinnissa, minkä vuoksi työssä ei käsitellä laajasti kehitystoimenpiteitä laadun parantamiseksi. Työssä keskitytään datan arvojen mittaamiseen ja arviointiin, mutta teoriassa tuodaan esille myös muita laadun näkökulmia kokonaisymmärryksen muodostamiseksi. Teoriassa käsitellään dataa yleisesti, mutta empiriassa tarkasteltava data on rajattu Master asiakasdatan yhteystietoihin.

Diplomityön päätutkimuskysymys on:

- *Miten kohdeyrityksen datan laatua voidaan mitata ja arvioida?*

Työn alatutkimuskysymykset ovat:

- *Mitä datan laadulla tarkoitetaan?*
- *Minkälaisia menetelmiä datan laadun mittaamiseen ja arviointiin on olemassa?*
- *Miten data, laatu, mittaaminen ja arviointi liittyvät toisiinsa?*

Työn tarkoituksena on ensin vastata alatutkimuskysymyksiin ja löytää niiden sekä empirian avulla vastaus päätutkimuskysymykseen. Ensimmäiseen alatutkimuskysymykseen vastataan luvussa kaksi, toiseen alatutkimuskysymykseen vastataan luvuissa kolme ja neljä, ja viimeisen kokoavan alatutkimuskysymyksen vastausaineistona toimivat kaikki teorialuvut. Diplomityön tarkoituksena on muodostaa yleinen viitekehys datan laadun mittaamiseen ja arviointiin, mitä voidaan hyödyntää kohdeyrityksen eri datoihin.

1.3 Tutkimusvalinnat

Tutkimusvalintoja tarkastellaan tieteenfilosofian, lähestymistavan, strategian, menetelmän, aikahorisontin sekä aineiston keräämisen ja analysoinnin näkökulmista. Tutkimusvalinnat on esitetty taulukossa 1.

Taulukko 1. *Tutkimusvalinnat.*

Tutkimuksen näkökulmat	**Tutkimusvalinnat**
Tieteenfilosofia	Pragmatismi
Lähestymistapa	Deduktio (teorialähtöinen)
Strategia	Tapaustutkimus
Menetelmä	Yhdistelmämenetelmä (triangulaatio)
Aikahorisontti	Poikittaistutkimus (tietyn hetken poikkileikkauskuva)
Aineistojen kerääminen	Haastattelut (laadullinen ja määrällinen aineisto) sekä erillinen määrällinen aineisto kohdeyrityksen tietokannasta
Aineistojen analysointi	Sisällönanalyysi (luokittelu), toistuvuuden laskeminen, objektiiviset mittarit, roolien etäisyyksien analysointi ja vertailuanalyysi

Tutkimuksen tieteenfilosofiana on pragmatismi, jossa hyödynnetään eri näkökulmien yhdistämistä tietojen keräämisessä ja tulkinnassa (Saunders 2009 s. 598). Pragmatismissa tietoa voidaan kerätä yhdistettyjen tai monimetodisten asetelmien avulla (Saunders 2009 s. 119). Tutkimuksen lähestymistapa on puolestaan deduktiivinen eli teorialähtöinen, jossa kirjallisuudesta luodaan teoreettinen viitekehys. Deduktiivisen lähestymistavan ominaisuuksiin kuuluvat esimerkiksi käsitteiden operationalisointi siten, että tosiasioita

voidaan mitata määrällisesti sekä yleistäminen. (Saunders 2009 ss. 125-127) Tutkimuksen teoria on muodostettu datan laadun mittaamisen ja arvioinnin keskeisistä tieteellisistä artikkeleista ja kirjoista. Kirjallisuuskatsauksen aineisto on haettu pääosin hakulausekkeella *("data quality measurement") OR ("data quality metrics") OR ("data quality assessment")* Tampereen yliopiston Andor-tietokannasta ja Google Scholar-hakupalvelusta. Tässä työssä käytetään datan laadun mittaamisen ja arvioinnin teoriaa kohdeyrityksen datan laadun tason diagnosoimiseen ja valitun arviointimenetelmän soveltuvuuden testaamiseen.

Tutkimus toteutetaan tapaustutkimuksena. Tapaustutkimukseen liittyy empiirinen tutkimus tietystä nykyisestä ilmiöstä sen tosielämän asiayhteydessä (Saunders 2009 s. 145). Tapaustutkimuksessa tarkastellaan intensiivisesti yhtä tapausta tai pientä joukkoa toisiinsa suhteessa olevia tapauksia (Hirsjärvi et al. 2007). Siinä pyritään vastaamaan tutkimuskysymyksiin, jotka ovat muodoltaan ”Miten?” - ja ”Miksi?”- alkuisia (Yin 2003 ss. 5-7). Tässä tutkimuksessa tarkastellaan yhtä tapausta, kohdeyrityksen datan laadun mittaamisen ja arvioinnin käytäntöjen kehittämistä. Tutkimuksen aikahorisontiksi valittiin poikittaistutkimus, jossa tarkastellaan tietyn hetken poikkileikkauskuvaa (Saunders 2009 s. 155). Se valittiin, koska tutkimuksen tarkoituksena on selvittää kohdeyrityksen datan laadun taso tutkimuksen toteuttamishetkellä.

Tutkimuksessa käytetään yhdistelmämenetelmää, joka viittaa määrällisten ja laadullisten tiedonkeruu- ja analysointimenetelmien hyödyntämiseen joko samanaikaisesti tai peräkkäin. Voidaan puhua myös triangulaatiosta, jolla tarkoitetaan kahden tai useamman tiedonkeruumenetelmän käyttämistä tutkimuksen tuloksien vahvistamiseksi. (Saunders 2009 ss. 152-154) Tässä tutkimuksessa puolistrukturoitujen haastatteluiden avulla kerättyä laadullista ja määrällistä aineistoa trianguloidaan kohdeyrityksen tietokannasta saatavalla määrällisellä Master asiakasdatalla. Yhdistämisen tarkoituksena on selvittää mahdolliset yhteneväisyydet ja eroavaisuudet subjektiivisen haastatteluaineiston sekä objektiivisen Master asiakasdata-aineiston tuloksien välillä.

Haastattelut olivat puolistrukturoituja. Puolistrukturoiduissa haastatteluissa on mahdollista esittää haastattelurungosta poikkeavia lisäkysymyksiä (Saunders 2009 s. 320). Haastatteluissa hyödynnettiin lisäkysymyksiä, kun haluttiin tarkentaa ja selventää haastateltavien vastauksia. Haastattelut toteutettiin yksilöhaastatteluina Skypen välityksellä ja ne nauhoitettiin. Haastateltavien valinnassa hyödynnettiin harkinnanvaraista otosta, jossa valitaan ne henkilöt, jotka osaavat parhaiten vastata tutkimuskysymyksiin (Saunders 2009 s. 237). Harkinnanvaraista otosta käytettiin, koska haluttiin tarkastella tiettyjen työntekijäryhmien mielipidettä aiheen asiantuntemuksen perusteella. Haastatteluista saatiin laadullisen aineiston lisäksi myös määrällistä aineistoa, kun haastateltavat antoivat numeerisen arvon laadun eri ulottuvuuksille. Haastatteluiden analysoinnissa hyödynnettiin sisällönanalyysiä, jossa tutkimusaineistoa kuvataan sanallisesti ja tuodaan esiin merkityssuhteita sekä merkityskokonaisuuksia (Vilkka 2015). Lisäksi hyödynnettiin luokittelua,

kun samankaltaisia vastauksia sisällytettiin samoihin luokkiin ja laskettiin niiden esiintymiskerrat (Tuomi & Sarajärvi 2009 s. 93). Numeeristen haastatteluvastausten analysoinnissa puolestaan hyödynnettiin aritmeettisten keskiarvojen laskemista ja visualisointeja. Eri työntekijäryhmien tuloksien välillä oli nähtävissä eroavaisuuksia, minkä myötä hyödynnettiin roolien etäisyyksien analysointia.

Määrällisen Master data-aineiston valinnassa hyödynnettiin ryväsotantaa. Ryväsotanta on toimiva menetelmä, kun tutkimuskohteena ovat luonnolliset ryhmät, kuten yritykset, organisaatiot tai kaupunginosat. Ryppäät voidaan valita satunnaisesti tai systemaattisesti, ja valituille ryppäille voidaan tehdä kokonaistutkimus. (Vilkka 2015) Master asiakasdata voidaan jakaa ryppäisiin asuinalueittain, ja näistä ryppäistä valittiin systemaattisesti tietyn maakunnan keskustaajaman asiakkaat datan käsittelyn helpottamiseksi. Määrällisen Master data-aineiston analysoinnissa hyödynnettiin puolestaan objektiivisia mittareita. Lopuksi toteutettiin vertailuanalyysi haastatteluaineiston ja objektiivisten mittareiden välillä. Tulosten analysointimenetelmistä kerrotaan tarkemmin luvussa 5.4.

1.4 Tutkimuksen rakenne

Tutkimus koostuu kahdeksasta luvusta, joista kolmessa luvussa esitetään työn keskeinen teoria. Johdannon jälkeisessä ensimmäisessä teorialuvussa käsitellään datan laatua datan tyyppien ja luokitteluiden, laadun määritelmien ja näkökulmien, laadun ulottuvuuksien sekä laadun kustannuksien avulla.

Kolmannessa luvussa esitetään datan laadun mittaamisen teoria. Mittaamiseen pureudutaan ensin tutustumalla erilaisiin mittaustyyppeihin, jonka jälkeen käsitellään mittaamisen yleisiä vaatimuksia ja tarkempia vaatimuksia itse mittareille. Näiden jälkeen tutustutaan datan profilointiin ja datan laadun mittareihin. Lopuksi keskitytään datan laadun mittareiden esitystapoihin.

Neljännessä luvussa esitetään datan laadun arvioinnin teoria. Luvun painopiste on neljän erilaisen datan laadun arviointimenetelmän käsittelemisessä, joiden lisäksi tuodaan myös esille eri arviointimenetelmien vertailuperiaatteita. Lopuksi esitetään teoriaosuuden yhteenveto, jota voidaan pitää datan laadun mittaamisen ja arvioinnin viitekehyksenä.

Viidennessä luvussa esitetään työssä käytetyt empiiriset menetelmät. Työn empiria pohjautuu Hybridi-arviointimenetelmään, jonka vaiheiden mukaan viides luku on jaoteltu. Lisäksi esitetään tiiviisti myös kohdeyrityksen ominaispiirteitä. Työssä käytetään työntekijähaastatteluiden avulla saatua laadullista ja määrällistä aineistoa sekä kohdeyrityksen tietojärjestelmästä saatavaa määrällistä Master asiakasdataa.

Kuudennessa luvussa käsitellään tutkimuksen tulokset. Tuloksien esittäminen on jaoteltu subjektiivisten haastatteluiden tuloksien ja objektiivisten mittareiden tuloksien sekä näiden vertailuanalyysin tuloksien esittämiseen.

Seitsemännessä luvussa toteutetaan teorian ja empirian vertailua päätelmien muodostamiseksi. Kyseisessä luvussa pureudutaan tarkemmin tulosten päätelmiin, tutkimuskysymysten vastauksiin ja kehitysehdotusten esittämiseen.

Kahdeksannessa luvussa keskitytään tutkimuksen pohdintaan. Tiiviin tutkimuksen yleisen pohdinnan lisäksi esitetään myös tutkimuksen arviointia, rajoituksia ja potentiaalisia jatkotutkimusehdotuksia.

Liitteessä A on esitetty empiriassa hyödynnetty haastattelurunko, joka on muodostettu liitteessä B esitetyn AIMQ-kyselylomakkeen pohjalta.

2. DATAN LAATU

Yritysten kilpailun perusta on muuttunut aineellisista tuotteista aineettomiin tietoihin. Yritysten tiedot edustavat kollektiivista tietoa, jota käytetään tuottamaan ja toimittamaan tuotteita ja palveluita kuluttajille. Tiedon laatu tunnistetaan yhä useammin yrityksen arvokkaimmaksi eduksi. (McGilvray 2008 s. 4) Tieto on kuitenkin hyvin monitasoinen käsite, koska sillä voidaan tarkoittaa esimerkiksi dataa, informaatiota tai tietämystä.

Tuotteen laatu riippuu prosessien laadusta, miten tuote on suunniteltu ja tuotettu. Samoin datan laatu on riippuvainen niiden prosessien suunnittelusta ja tuottamisesta, mitkä liittyvät datan luomiseen. Paremman laadun tavoittelussa on ensin ymmärrettävä, mitä laatu tarkoittaa ja miten sitä mitataan. (Wand & Wang 1996)

Kirjallisuudessa on useita lähestymistapoja, joita voidaan soveltaa datan laadun tutkimiseen. Yksi niistä on datan elinkaari, mikä keskittyy toimintoihin datan luomisesta sen jakamiseen. (Wang et al. 1995) Datan laatua voidaan hallita myös tietojärjestelmien eri toimintojen näkökulmista. Datan laatu voi esimerkiksi koskea tietokantojen suunnittelua, kuten loogisen tai fyysisen tietokantamallin laatua. Datan laatu voi myös viitata datan arvoihin, joita lisätään, tallennetaan ja päivitetään tietovirran aikana. (Boyadzhieva & Kolev 2010) Tässä työssä keskitytään datan arvojen laatuun, mutta tuodaan esiin myös muita datan laadun näkökulmia kokonaisymmärryksen muodostamiseksi.

2.1 Datan luokittelu

Datan määrittelemisessä voidaan hyödyntää tiedon tasojen näkökulmaa datasta, informaatiosta ja tietämyksestä. Datalla tarkoitetaan rakenteettomia tosiasioita, informaatio puolestaan viittaa analyyseissä hyödynnettävään rakenteelliseen dataan ja tietämys on kokemukseen perustuvaa inhimillistä tietoa. Datasta voidaan luoda informaatiota luomalla sille rakenne ja informaatiota tulkittaessa saadaan tietämystä. (Laihonen et al. 2013) Data- ja informaatio- termejä käytetään usein synonyymisesti. Käytännössä ne kuitenkin eroavat toisistaan, sillä informaatiolla tarkoitetaan prosessoitua dataa. (Pipino et al. 2002)

Data erotetaan usein kolmeen eri tyyppiin:

1. Rakenteelliset datat (*engl. Structured data*) ovat koosteita tai yleistyksiä asioista. Relaatiotaulut ja tilastollinen data edustavat yleisintä rakenteellisen datan tyyppiä. (Batini et al. 2009)
2. Rakenteettomat datat (*engl. Unstructured data*) ovat yleisiä symbolien sarjoja, jotka ovat tavallisesti koodattu luonnollisella kielellä (Batini et al. 2009). Rakenteetonta dataa ei voi tallentaa riveinä ja kolumneina relaatiotietokantaan. Esimerkiksi kuvat ja videot edustavat rakenteetonta dataa. (Aljumaili et al. 2016)

3. Puolirakenteellisilla datoilla (*engl. Semistructured data*) on puolestaan jonkinasteista joustavuutta. Puolirakenteellista dataa kutsutaan myös skeemattomaksi tai itseään kuvailevaksi. (Batini et al. 2009) Ne edustavat osittain rakenteellista dataa, mutta niillä ei ole tarkkaa datamallin rakennetta (Aljumaili et al. 2016).

Datan laadun kirjallisuudessa keskitytään pääosin rakenteelliseen dataan. Yksi syy tähän on se, että kyseistä dataresurssia hyödynnetään eniten useimmissa organisaatioissa. (Batini et al. 2009)

Fyysisten tuotteiden ja datatuotteiden valmistamisen välillä voidaan nähdä samankaltaisuuksia. Tuotteiden valmistusjärjestelmä hyödyntää raaka-aineita tuottaakseen fyysisiä tuotteita. Samankaltaisesti tietojärjestelmä voidaan nähdä datan valmistusjärjestelemänä, jossa hyödynnetään raakaa dataa (esim. yksittäisiä numeroita, tietueita, tiedostoja, laskentataulukkoja tai raportteja) tuottamaan dataa tai datatuotteita, kuten lajiteltuja tiedostoja tai korjattuja postituslistoja. Tätä datatuotetta puolestaan voidaan käsitellä raakana datana toisessa datan valmistusjärjestelmässä. (Wang et al. 1995; Ballou et al. 1998) Taulukossa 2 on esitetty analogia fyysisten tuotteiden ja datatuotteiden valmistamisen välillä.

Taulukko 2. *Fyysisten tuotteiden ja datatuotteiden valmistaminen (mukaillen Wang et al. 1995)*

	Tuotteiden valmistaminen	**Tiedon valmistaminen**
Sisääntulo	Raakamateriaalit	Raakadata
Prosessi	Materiaalien prosessointi	Datan prosessointi
Ulostulo	Fyysiset tuotteet	Datatuotteet

Termi ”datan valmistaminen” kannustaa etsimään poikkitieteellisiä analogioita, jotka voivat helpottaa tietämyksen siirtämisessä tuotteiden laadun alalta datan laadun alalle. Termi ”datatuote” puolestaan korostaa datatuotoksen arvoa, joka siirretään asiakkaille. (Wang et al. 1995)

Dataa voidaan luokitella niiden yhteisten ominaispiirteiden mukaan. Luokittelut ovat hyödyllisiä datan hallinnan kannalta, koska tiettyjä datoja saatetaan kohdella luokittelun perusteella eri tavalla. Suhteiden ja riippuvuuksien ymmärtäminen eri kategorioiden välillä voi auttaa datan laadun ohjaamisessa. (McGilvray 2008 s. 39**)** Taulukossa 3 on esitetty yleisimmät datan kategoriat ja niiden määritelmät.

Taulukko 3. *Datan kategoriat ja niiden määritelmät (mukaillen McGilvray 2008 ss. 42-43)*

Datan kategoria	Määritelmä
Master data	Master data kuvaa organisaation liiketoimintaan liittyviä ihmisiä (esim. asiakkaat), paikkoja (esim. myyntialueet) ja asioita (esim. tuotteet).
Transaktiodata	Transaktiodata kuvaa sisäistä tai ulkoista tapahtumaa tai tapahtumaa, joka ilmentyy liiketoiminnan harjoittamisen myötä. Esimerkkejä ovat myyntitilaukset, laskut ja tilaukset.
Referenssidata	Referenssidataa ovat arvojoukot tai luokittelumalit, mihin viitataan esim. järjestelmissä, tietovarastoissa ja prosesseissa. Esimerkkejä ovat validien arvojen luettelot, koodilistat, valtion lyhenteet ja tuotetyypit.
Metadata	Metadatalla tarkoitetaan ”dataa datasta”. Metadata kuvailee muita datoja, mikä tekee datan hakemisesta, tulkitsemisesta ja käyttämisestä helpompaa. Se voidaan jakaa tekniseen (esim. kenttien pituudet ja tyypit), liiketoiminnalliseen (esim. kenttien määritelmät) ja jäljitysketjuun liittyvään metadataan (esim. datan päivittäjän nimi ja tunnus).
Historiadata	Historiadata sisältää tietyn ajankohdan merkittäviä tosiasioita, jotka ovat tärkeitä turvallisuuden ja ohjeidenmukaisuuden kannalta. Esimerkkejä ovat tietokannan tilannekatsaus ja versiotiedot.
Väliaikainen data	Väliaikaista dataa säilytetään muistissa prosessin nopeuttamiseksi ja niitä käytetään vain teknisiin tarkoituksiin. Esimerkkinä voidaan pitää taulukon kopiota, joka luodaan nopeuttamaan hakuja.

Dataa voidaan myös luokitella eri tavoin kuin taulukossa 3 on kuvattu. Voi olla vaikea päättää, onko esimerkiksi validien arvojen lista vain referenssidataa vai myös metadataa. Referenssidataa tarvitaan Master datan luomiseksi ja Master dataa tarvitaan transaktiodatan luomiseksi. Joskus referenssidataa voidaan tarvita myös transaktiodatan luomiseen. Metadataa puolestaan tarvitaan muiden datan kategorioiden ymmärtämiseen. (McGilvray 2008 ss. 43-44)

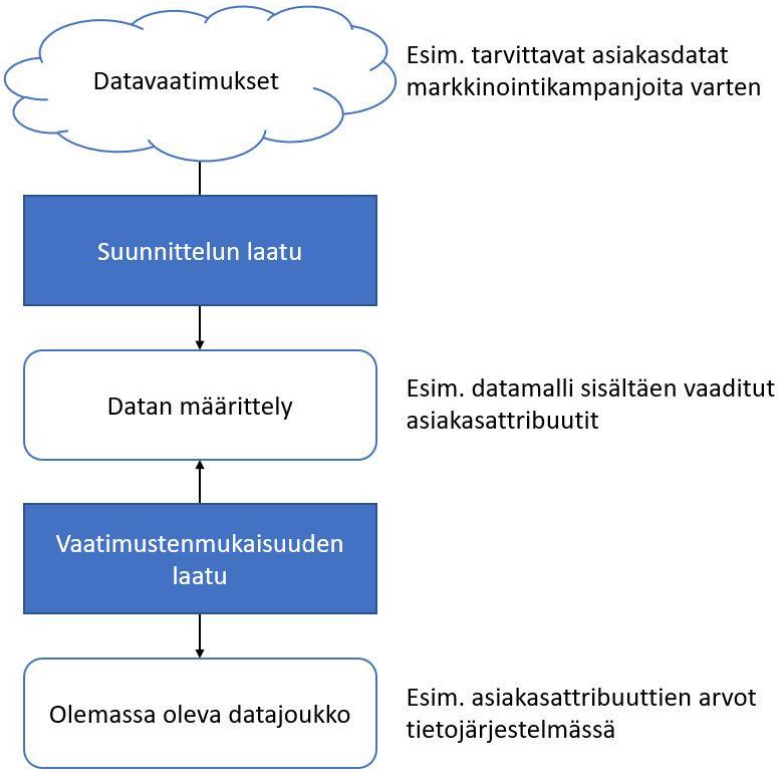
2.2 Laadun näkökulmat

Määritelmä “sopivuus käyttötarkoitukseen” (*engl. Fitness for use*) on laajasti hyväksytty laatukirjallisuudessa. Se korostaa kuluttajan näkökulman tärkeyttä laatuun, koska lopulta kuluttaja arvioi tuotteen sopivuutta käyttöön. (Wang & Strong 1996) Tähän viittaa myös Umar et al. (1999) määritelmässään, jonka mukaan tuote, palvelu tai tieto X on laadultaan korkeampi kuin tuote, palvelu tai tieto Y, jos X täyttää asiakkaiden tarpeet paremmin kuin

Y. Sebastian-Coleman (2013 s. 40) esittää, että datan laadun taso kuvaa sitä, missä määrin data vastaa datan kuluttajien odotuksia. Datan laatu liittyy siis suoraan datan käyttötarkoituksiin. Yksi keskeinen tekijä tämän ymmärtämisessä on se, miten hyvin data esittää sen kuluttajien mielestä sitä, mitä sen on tarkoituskin esittää.

Datan laatu voidaan määritellä myös tietojärjestelmän datanäkymien ja reaalimaailman datan yhdenmukaisuuden mittana (Heinrich et al. 2018b). Datan laatu antaa tietoa siitä, kuinka laajasti dataa puuttuu tai on virheellistä. Datan laatu voidaan myös määritellä keskittymällä tehtävän prosessin luonteeseen: korkealaatuinen data sopii tarkoitettuihin käyttötarkoituksiin, kuten päätöksentekoon, suunnitteluun ja tuotannon järjestelmiin. (Boyadzhieva & Kolev 2010)

Laatua voidaan tarkastella kahdesta erilaisesta laatua koskevasta konseptista ja määritelmästä, jotka vaikuttavat myös laadun määrittämiseen: suunnittelun laatu (*engl. Quality of design*) ja vaatimustenmukaisuuden laatu (*engl. Quality of conformance*). Suunnittelun laatu tarkoittaa käyttäjien vaatimusten ja tietojärjestelmän määritelmän vastaavuuden astetta. Vaatimuksenmukaisuuden laatu puolestaan edustaa tietojärjestelmien määrittelyn ja olemassa olevan toteutuksen vastaavuuden astetta, esimerkiksi datamallia verrataan tallennettuihin datan arvoihin. (Heinrich et al. 2009) Kuvassa 1 on havainnollistettu suunnittelun laatua ja vaatimustenmukaisuuden laatua.



Kuva 1. *Suunnittelun laatu ja vaatimustenmukaisuuden laatu (mukaillen Heinrich et al. 2009)*

Suunnittelun laadun ja vaatimustenmukaisuuden laadun erottaminen on tärkeää myös datan laadun määrittämisen kontekstissa. Se jakaa enimmäkseen subjektiiviset analyysit käyttäjien vaatimusten ja datamallien määritelmien vastaavuuden välillä sekä objektiivisemmat vastaavuuden määrittelyt datamallin ja olemassa olevien datan arvojen välillä. Vaatimuksenmukaisuuden laadun mittareita voidaan soveltaa monissa eri tilanteissa ja ne ovat uudelleenkäytettävämpiä, koska ne ovat riippumattomia tiettyjen käyttäjien vaatimuksista tietyssä liiketoimintaympäristössä. (Heinrich et al. 2009)

Data, joka aiemmin täytti yrityksen tietyn toiminnallisen alueen tarpeet, yhdistetään nyt myös muihin toimintoalueisiin. Samalle datalle on erilaisia liiketoiminnallisia käyttötarkoituksia, erilaisia alustoja, järjestelmiä, tietokantoja, sovelluksia, erityyppisiä datoja sekä erilaisia datarakenteita, määritelmiä ja standardeja. Dataa, prosesseja ja teknologiaa mukautetaan myös liiketoiminnan, maantieteellisen sijainnin tai sovelluksen mukaan. Näitä voidaan pitää nykyisen ympäristön haasteina. (McGilvray 2008 s. 6)

Datan laatuongelmia voi ilmentyä esimerkiksi datan hankinnan, tallentamisen, jakamisen ja ylläpidon aikana (Liu et al. 2018). Laatuongelmat voivat johtua ihmisten, prosessien tai järjestelmien ongelmista. Yrityksissä voidaan olla tietoisia, että data saattaa aiheuttaa aika ajoin ongelmia. Voi olla kuitenkin vaikea havaita, missä määrin nämä ongelmat vaikuttavat liiketoimintaan. (McGilvray 2008 s. 5) Laatuongelmia voidaan luokitella esimerkiksi seuraavasti:

- Datan näkymiin liittyvät ongelmat, kuten datan tärkeys ja yksityiskohtaisuus.
- Datan arvoihin liittyvät ongelmat, kuten datan tarkkuus, johdonmukaisuus, ajantasaisuus ja täydellisyys.
- Datan esittämiseen liittyvät ongelmat, kuten datan formaatin tarkoituksenmukaisuus ja tulkinnan helppous.
- Muut ongelmat, kuten yksityisyys, turvallisuus ja omistajuus. (Redman 1998)

Suuret ja monimutkaiset järjestelmät sisältävät useita komponentteja, kuten esimerkiksi dataa, sitä hyödyntävän ohjelmiston, taustalla olevia alustoja sekä prosessin järjestelmän käyttämiseen ja hallitsemiseen. Eri toimijat (esim. järjestelmän käyttäjät, johtajat, yritysasiakkaat) katsovat näitä komponentteja eri tasoilla. Käsitteellisesti erilaisia datan laadun näkemyksiä voidaan ilmaista kunkin komponentin suhteen osoittaakseen sen käyttäytymistä ja parantaakseen sen laatua. Datan laadun kannalta on tärkeää ottaa huomioon myös monia suoraan dataan liittymättömiä ongelmia, kuten alustoja, prosesseja ja ohjelmistoarkkitehtuuria. Esimerkkejä näkemyksistä ovat:

- Itse datan laatu (esim. tarkkuus, ajantasaisuus, yhdenmukaisuus, täydellisyys).
- Ohjelmiston laatu (esim. mahdolliset ohjelmistovirheet).
- Alustan laatu (esim. käyttötapausten suorituskyky).
- Hallinta- ja toimintaprosessien laatu (esim. virheet, viivästykset, läpivirtaukset, käyttäjien tyytyväisyysaste). (Umar et al. 1999)

Datan laadunhallinta tarkoittaa erilaisten datan laatuongelmien tunnistamista, mittaamista ja seurantaa. Datan laadunhallinnan toimintoja kehittämällä voidaan parantaa datan laatua. (Liu et al. 2018) Datan laadun parantamiseksi on tehtävä useita arviointi- ja parannustoimia sen koko elinkaaren ajan. Datan laadun kehittämisen lähestymistavat voidaan jakaa kahteen laajaan luokkaan, jotka ovat datan siivoaminen ja prosessien siivoaminen. On kuitenkin hyvä huomioida, että kokonaisvaltaiseen laadunhallintaan tarvitaan näitä molempia näkökulmia. Datan siivoaminen edellyttää työkalun käyttämistä huonolaatuisen datan (esimerkiksi epätarkan, -yhdenmukaisen, -ajankohtaisen tai -täydellisen) tunnistamiseen ja siten huonojen datojen poistamiseen automaattisten tai manuaalisten prosessien avulla. Tämän lähestymistavan päärajoituksena on, että kaikkia tietoja ei voida helposti todentaa oikeiksi. Lisäksi datan puhdistamisen on oltava säännöllistä koko datan elinkaaren ajan. (Umar et al. 1999)

Prosessien siivoaminen menee datan siivoamisen taustalle ja keskittyy toimintoihin, jotka heikentävät hyvälaatuista dataa. Sen keskeisimmät toiminnot ovat laatumittarien luominen, datan elinkaaren seuraaminen laadun saastumien varalta sekä tilastollisen laadunvalvonnan ja prosessienhallinnan käyttäminen halutun datan laadun ylläpitämiseksi. (Umar et al. 1999)

Datan poikkeavuuksia voi ilmentyä kaikissa datan elinkaaren vaiheissa, joten korkealaatuisen datan saamiseksi on asetettava useita datan laadun tarkistuksia järjestelmään. Tämän lisäksi on myös sovellettava erilaisia menetelmiä datan laatuongelmiin huomioimalla niiden alhainen datan laatu tai myös tekemällä korjauksia. Jos taas joitain dataan liittyviä ongelmia ei voida korjata, niin niitä ei kuitenkaan tulisi sivuuttaa, vaan tallentaa ja huomioida alhainen laatu niiden nimeämisessä. (Boyadzhieva & Kolev 2010)

Monet lähestymistavat pyrkivät tunnistamaan ja siivoamaan integraatioprosessin aikana syntyneitä virheitä datassa. Tarkoituksena on, että vain korkealaatuista dataa syötetään tietokantaan tai tietovarastoon, mutta syötetyn datan laatua ei kuitenkaan mitata tarkasti. Datan laatu yleensä myös heikkenee ajan kuluessa, mikä tuo haasteita laadun tason ylläpitämiseen. (Boyadzhieva & Kolev 2010)

Datan laadulle voidaan esittää myös alustava käsitteellinen kehys, joka sisältää seuraavat näkökohdat:

- Datan on oltava kuluttajien käytettävissä. Esimerkiksi kuluttaja tietää, miten dataa haetaan.
- Kuluttajan on kyettävä tulkitsemaan dataa. Esimerkiksi dataa ei esitetä vieraalla kielellä.
- Datan on oltava merkityksellistä kuluttajalle. Esimerkiksi data on asiaankuuluvaa ja ajankohtaisesti käytettävissä päätöksentekoprosessissa.
- Datan on oltava tarkkaa. Esimerkiksi data on virheetöntä, objektiivista ja se tulee hyvämaineisista lähteistä. (Wang & Strong 1996)

Datan laatu voidaan määritellä datana, joka sopii käyttötarkoitukseen. Datan kuluttajille sopivuus käyttötarkoitukseen tarkoittaa, että data on tarkkaa, uskottavaa, objektiivista, merkityksellistä, ajankohtaista, hyvämaineista, lisäarvoa tuovaa, tiiviisti ja johdonmukaisesti esitetty, täydellistä, tulkittavaa, ymmärrettävää, helposti saatavissa, turvallista sekä dataa on sopiva määrä. Tyypillisesti datan laadun kehittämisprojekteissa näistä ulottuvuuksista valitaan osajoukko. (Yang et al. 2004)

2.3 Laadun ulottuvuudet

Laadun ulottuvuudet ovat joukko datan laatuattribuutteja, jotka esittävät yhtä datan laadun näkökulmaa tai rakennetta (Wang & Strong 1996). Ulottuvuudet tarjoavat tavan datan laadun mittaamiseen ja hallintaan (McGilvray 2008 s. 30). Laadun ulottuvuuksilla voidaan viitata esimerkiksi datan arvoihin tai niiden malleihin, mutta useimmat määritelmät datan laadun ulottuvuuksista ja mittareista viittaavat kuitenkin datan arvoihin (Batini et al. 2009).

Datan laadun ulottuvuus on yleisesti mitattava kategoria tietyn datan ominaisuuden mukaan. Laadun ulottuvuudet mahdollistavat laadun ymmärtämisen suhteessa mittakaavaan ja muihin saman mittakaavan mittauksiin tai eri mittakaavoihin, joiden suhde on määritelty. Datan laadun ulottuvuuksien joukkoa voidaan käyttää odotuksien määrittämisessä halutulle datalle sekä datan laadun tilan mittaamisessa. (Sebastian-Coleman 2013 s. 40)

Monille ihmisille datan laatu tarkoittaa vain tarkkuutta. Datan laatu on kuitenkin paremmin edustettuna, jos sitä mitataan myös muiden laadullisten ominaispiirteiden mukaan. Mitattavien laadun ulottuuksien valinta riippuu käyttäjien vaatimuksista. (Boyadzhieva & Kolev 2010) Taulukossa 4 on esitetty keskeisimpiä kirjallisuudessa esiintyviä laadun ulottuvuuksia ja niiden määritelmiä. Numerot viittaavat lähteeseen, jossa kyseinen ulottuvuus esiintyy.

Taulukko 4. *Laadun ulottuvuuksia ja niiden määritelmiä (mukaillen 1=Wang & Strong 1996; 2=Pipino et al. 2002; 3=Lee et al. 2002; 4=Batini et al. 2009; 5=Sebastian-Coleman 2013 ss. 62-63; 6=Heinrich et al. 2018a)*

Ulottuvuus	Määritelmä	1	2	3	4	5	6
Ajantasaisuus	Data ei ole vanhaa				X		
Asiaankuuluvuus	Data on olennaista tehtävää varten	X	X	X			
Eheys	Data noudattaa datamallin suhdesääntöjä					X	
Esityksen johdonmukaisuus	Data esitetään samassa muodossa	X	X	X			
Esityksen ytimekkyys	Data on tiiviisti esitetty	X	X	X			

Helppokäyttöisyys	Dataa voidaan käyttää eri käyttötarkoituksiin		X	X			
Jalostusarvo	Datan käyttämisestä saadaan hyötyä	X	X				
Johdonmukaisuus	Data on johdonmukaista sääntöjen, standardien tai muun datan suhteen				X	X	X
Maine	Datalla ja datalähteellä on hyvä maine	X	X	X			
Objektiivisuus	Data on tasapuolista ja ennakkoluulotonta	X	X	X			
Oikea-aikaisuus	Data on oikeaan aikaan kuluttajien käytettävissä tai ajantasaista tehtävää varten	X	X	X	X	X	X
Oikeellisuus	Data vastaa reaalimaailman arvoja						X
Saatavuus	Data on kuluttajien käytettävissä	X	X	X			
Sopiva määrä	Dataa on sopiva määrä tehtävää varten	X	X	X			
Tarkkuus	Data on oikeaa ja virheetöntä	X			X		
Tulkittavuus	Datan kielet, symbolit, yksiköt ja määritelmät ovat selkeitä	X	X	X			
Turvallisuus	Datan saatavuutta voidaan rajoittaa	X	X	X			
Täydellisyys	Datassa on kaikki tarvittavat osat	X	X	X	X	X	X
Uskottavuus	Data on todenmukaista ja luotettavaa	X	X	X			X
Validius	Data noudattaa liiketoimintasääntöjä					X	
Virheettömyys	Data on virheetöntä ja tarkkaa		X	X			
Volatiliteetti	Data on ajallisesti validia		X		X	X	
Ymmärrettävyys	Data on helposti ymmärrettävissä	X	X	X			

Ei ole kuitenkaan yleistä yhteisymmärrystä datan laadun ulottuvuuksien tarkoista merkityksistä tai siitä, mitkä ulottuvuudet määrittelevät datan laadun. Eroavaisuudet laadun ulottuvuuksien määritelmissä johtuvat etenkin laadun kontekstuaalisesta luonteesta. (Batini et al. 2009) Esimerkiksi johdonmukaisuutta (*engl. Consistency*) voidaan tarkastella esitystavan, sääntöjen, standardien tai muun datan suhteen.

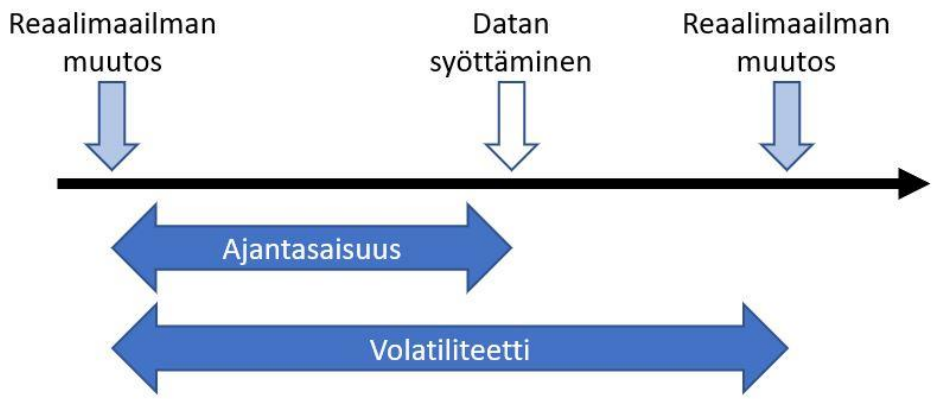
Laadun ulottuvuuksia voidaan luokitella neljään eri kategoriaan; luontaiseen laatuun, kontekstuaaliseen laatuun, esitystavan laatuun ja saavutettavuuden laatuun (Wang &

Strong 1996; Lee et al. 2002). Nämä kategoriat ja niihin liittyvät ulottuvuudet ovat esitetty kuvassa 2.

Kuva 2. *Datan laadun ulottuvuuksien kategoriat (mukaillen Wang & Strong 1996; Lee et al. 2002)*

Luontainen laatu viittaa datan itsenäiseen laatuun. Kontekstuaalinen datan laatu korostaa vaatimusta, jonka mukaan datan laatua tulisi tarkastella tehtävän kontekstin yhteydessä. Esitystavan ja saavutettavuuden laatu painottavat järjestelmien roolin tärkeyttä, esimerkiksi järjestelmän on oltava turvallisesti käytettävissä ja sen on esitettävä dataa laadukkaasti. Näiden kategorioiden perusteella voidaan todeta, että korkealaatuinen data on luontaisesti laadukasta, asiayhteyteen sopivaa, selkeästi esitetty ja saatavissa datan kuluttajille. (Wang & Strong 1996; Lee et al. 2002)

Aikaan liittyviä ulottuvuuksia ovat ajantasaisuus (*engl. Currency*), oikea-aikaisuus (*engl. Timeliness*) ja volatiliteetti (*engl. Volatility*). Kirjallisuudessa hyvin erilaisia määritelmiä ajantasaisuudelle ja oikea-aikaisuudelle, ja niillä voidaan viitata myös samaan käsitteeseen. (Batini et al. 2009) Oikea-aikaisuudella tarkoitetaan, kuinka ajantasaista data on tarkoitettuun tehtävään nähden. Ajantasaisuus puolestaan viittaa datayksiköiden ikään ja volatiliteetti tarkoittaa sitä ajan pituutta, jolloin data pysyy vielä validina. (Ballou et al. 1998; Pipino et al. 2002) Volatiliteetti voidaan myös määritellä reaalimaailman muutoksen ja alkuperäisen datan vääristävän myöhemmän muutoksen välisenä aikana. Oikea-aikaisuuden määrittämisessä voidaan hyödyntää kolmea tapahtumaa; ensimmäinen on reaalimaailman muutos, toinen on muutoksen tallentaminen datana tietojärjestelmään ja kolmas on tämän datan hyödyntäminen. (Blake & Mangiameli 2011) Kuvassa 3 on esitetty hahmotelma oikea-aikaisuuteen liittyvistä käsitteistä.



Kuva 3. *Tapahtumat, joita käytetään ajantasaisuuden ja volatiliteetin määrittämiseen (mukaillen Blake & Mangiameli 2011)*

Tarkkuudella (*engl. Accuracy*) viitataan datan oikeellisuuteen ja totuuteen, ja se on yksi haastavimmista ulottuvuuksista. Sen mittaaminen ei ole yksinkertaista, sillä se edellyttää datan vertaamista reaalimaailman arvoihin. (Sebastian-Coleman 2013 s. 63) Oikeellisuus on todennäköisesti kaikkein tärkein ulottuvuus. Data on oikeaa, jos se vastaa todellisuutta. (Bronselaer et al. 2018a) Virheellisen datan määrittämiseksi voidaan asettaa kriteerejä, mutta ei ole mahdollista määrittää oikeaa tai tarkkaa dataa, ellei sitä voida verrata täysin oikeelliseen dataan. Esimerkiksi, jos samaan henkilöön viittaavat tietueet osoittavat kahta eri syntymäaikaa, niin voidaan päätellä vähintään toisen niistä olevan virheellinen. Ei ole kuitenkaan mahdollista määrittää oikeaa reaalimaailman totuutta viittaamatta ulkopuoliseen lähteeseen tai standardiin, joka vahvistaa tämän tosiasian. (Sebastian-Coleman 2013 s. 63) Tarkkuuden ja oikeellisuuden todentamisessa voidaan hyödyntää referenssidataa oikeista arvoista. Kyseistä tekniikkaa käytetään usein osoitetietojen oikeellisuuden mittaamiseen. (Bronselaer et al. 2018a)

Validiteettiä (*engl. Validity*) voidaan mitata, mutta validiteetti ei kuitenkaan tarkoita tarkkuutta tai oikeellisuutta, sillä validit arvot voivat olla vääriä. Validiteetin mittaamisessa hyödynnetään reaalimaailman kohteiden korvikkeita tai vastikkeita, jotka voidaan tunnistaa dataksi. Validiteetin mittaamisesta voidaan saada ymmärrystä, sillä epävalidit arvot eivät voi olla oikeita. (Sebastian-Coleman 2013 ss. 63-64)

Jokainen datan laadun ulottuvuus vaatii erilaisia työkaluja, tekniikoita ja prosesseja sen mittaamiseksi. On tärkeää ymmärtää vaatimukset eri ulottuvuuksien arvioimiseen, minkä avulla voidaan valita tarpeisiin sopivat ulottuvuudet. Laadun ulottuvuuksien ymmärtäminen auttaa:

- Valitsemaan liiketoimintatarpeisiin soveltuvat ulottuvuudet ja priorisoimaan niitä.
- Ymmärtämään, mitä eri ulottuvuuksien mittaamisesta ja arvioinnista saadaan.
- Määrittelemään ja hallitsemaan projektien toimintaa aikataulu- ja resurssirajoituksissa. (McGilvray 2008 ss. 30-31)

2.4 Laadun kustannukset

Kustannukset ovat oleellinen näkökulma, johtuen huonolaatuisen datan vaikutuksista resurssien kuluttamiseen. Datan laadun kustannukset ovat laadun arvioinnin ja parannustoimien kustannusten summa, mitä kutsutaan myös datan laatuohjelman kustannuksiksi. Huonolaatuisen datan kustannuksia voidaan vähentää toteuttamalla entistä tehokkaampaa datan laatuohjelmaa, joka on tyypillisesti kalliimpaa. Tämän vuoksi datan laatuohjelman kustannuksien lisääminen vähentää huonolaatuisen datan kustannuksia. Tämä vähennys voidaan nähdä datan laatuohjelman hyötynä. (Batini et al. 2009)

Datan laatuohjelman kustannuksia voidaan pitää ehkäisevinä kustannuksina, joilla organisaatiot vähentävät datan virheitä. Tämä kustannusluokka sisältää kaikkien niiden vaiheiden kustannukset, mitkä muodostavat datan laadun arvioinnin ja kehittämisen prosessin. Huonolaatuisen datan kustannukset voidaan luokitella seuraavasti:

1. Prosessikustannukset, kuten koko prosessin uudelleensuorittamiseen liittyvät kustannukset.
2. Vaihtoehtokustannukset menetetyistä tuloista. (Batini et al. 2009)

Huonolaatuisen datan kustannukset ovat vahvasti riippuvaisia kontekstista, toisin kuin datan laatuohjelman kustannukset. Tämän vuoksi sen arviointi on vaikeaa, koska samalla datan arvolla ja vastaavalla laadun tasolla voi olla eri vaikutus vastaanottajasta riippuen. (Batini et al. 2009)

Eppler & Helfert (2004) puolestaan esittävät, että datan laadun kustannukset koostuvat kahdesta päätyypistä, huonon datan laadun aiheuttamista kustannuksista ja parannuskustannuksista. Parannuskustannukset voidaan luokitella datan laatuprosessin mukaan ennaltaehkäisy-, selvitys- ja korjauskustannuksiin. Huonolaatuisen datan aiheuttamat kustannukset voidaan luokitella niiden mitattavuuden tai vaikutuksen mukaan suoriin ja epäsuoriin kustannuksiin. Taulukossa 5 on esitetty datan laadun kustannuksien luokittelua.

Taulukko 5. *Datan laadun kustannukset (mukaillen Eppler & Helfert 2004)*

Datan laadun kustannukset		
Huonon datan aiheuttamat kustannukset	Suorat kustannukset	Vahvistuskustannukset
		Uudelleensyöttämisen kustannukset
		Korvauskustannukset
	Epäsuorat kustannukset	Alhaisen maineen kustannukset
		Väärien päätösten kustannukset
		Hukatut investointikustannukset
Datan laadun parantamisen kustannukset	Ehkäisykustannukset	Koulutuskustannukset
		Seurantakustannukset
		Kehittämisen kustannukset
	Selvityskustannukset	Analyysikustannukset
		Raportointikustannukset
	Korjauskustannukset	Korjauksien suunnittelukustannukset
		Korjauksien toteuttamiskustannukset

Suorat kustannukset aiheutuvat huonosta datan laadusta ja niillä on negatiivisia rahallisia vaikutuksia. Niitä ovat kyseenalaisen uskottavuuden omaavan datan vahvistamisesta aiheutuvat kustannukset, virheellisen tai epätäydellisen datan uudelleensyöttämisen kustannukset sekä muille aiheutuneiden vahingonkorvausten kustannukset. Epäsuorat kustannukset puolestaan aiheutuvat huonosta datan laadusta välillisten vaikutusten myötä. Niitä ovat hintapreemion menetykset maineen huonontumisen vuoksi, huonoon dataan pohjautuvien epäedullisten päätösten kustannukset sekä menetetyt investointikustannukset. Datan laadun parantamiseen liittyviä kustannuksia ovat puolestaan koulutuskustannukset datan laadun ymmärryksen lisäämiseksi, seurantakustannukset, kehittämisen kustannukset, analyysikustannukset, raportointikustannukset sekä korjauksien suunnittelu- ja toteuttamiskustannukset. (Eppler & Helfert 2004)

Huonon datan laadun vaikutuksiin kuuluvat myös asiakkaiden tyytymättömyys, lisääntyneet toimintakustannukset, tehottomampi päätöstenteko ja heikentynyt kyky toteuttaa strategiaa. Lisäksi se vähentää työntekijöiden moraalia, lisää organisaation epäluuloa ja vaikeuttaa yrityksen yhdensuuntaistamista. Johtavat yritykset ovat kuitenkin osoittaneet, että datan laatua voidaan merkittävästi parantaa. (Redman 1998)

3. DATAN LAADUN MITTAAMINEN

Nykyaikaisen näkemyksen datan laadusta ovat esittäneet Wang et al. (1995) ja Wang & Strong (1996). He ovat tuoneet esiin, että laatu on datan monimutkainen ominaisuus, jota ei voida mitata suoraan. Sen sijaan on otettava huomioon erilaiset laadun ulottuvuudet, jotka ovat merkityksellisiä tiettyyn sovelluskohteeseen ja kehitettävä mittausmenetelmät kyseisille ulottuvuuksille. Tämä ymmärrys on johtanut useisiin lähestymistapoihin datan laadun ulottuvuuksien mittaamiseksi ja arvioimiseksi. (Bronselaer et al. 2018b)

Datan kuluttajat arvioivat laatua tietyissä liiketoimintakonteksteissa tai päätöksentekotehtävissä. Samalla dataresurssilla voi olla hyväksyttävä laadun taso joissain asiayhteyksissä, mutta tämä laatu voi hyväksymiskelvoton muissa asiayhteyksissä. (Shankaranarayanan & Even 2007) Samalla kun tutkimukset ja käytännöt ovat huomanneet hyvin perusteltujen datan laatumittarien merkityksen, niin monista datan laatumittareista puuttuu kuitenkin asianmukaiset metodiset perustat. Useat mittarit kehitetään ad hoc -periaatteella tiettyjen ongelmien ratkaisemiseksi tai ne ovat hyvin subjektiivisia. (Pipino et al. 2002; Heinrich et al. 2009; Heinrich et al. 2018a) Kaikissa datan laadun mittaamisen menetelmissä on kriittistä määritellä laatu, ulottuvuudet ja mittarit. Yleensä useita mittareita voidaan yhdistää yhteen laadun ulottuvuuteen. (Batini et al. 2009)

Datan laadun terminologioiden eroavaisuuksien vuoksi on tärkeää korostaa mittaamisen ja arvioinnin ero. Caballero et al. (2007) määrittävät mittaamisen toiminnoksi, jossa määritetään numeroarvo tarkastelun kohteena olevalle attribuutille. Arvioinnissa puolestaan luokitellaan joku tai jokin sen arvon perusteella. Mittaamisessa käytetään kvantitatiivisia eli määrällisiä arvoja, kun taas arvioinnissa kvalitatiivisia eli laadullisia arvoja. Batini et al. (2009) tarkentavat, että mittaus-termiä käytetään datan laadun ulottuvuuksien arvon mittaamisen yhteydessä. Arviointi-termiä puolestaan käytetään silloin, kun kyseisiä mittauksia verrataan vertailuarvoihin laadun diagnosoinnin mahdollistamiseksi. Sebastian-Coleman (2013 s. 46) tuo esiin näkemyksen, jonka mukaan mittaaminen tarkoittaa jonkin koon, määrän tai asteen selvittämistä välineen avulla. Mittauksen synonyyminä arviointi puolestaan merkitsee tarvetta verrata asioita toisiinsa ymmärryksen luomiseksi.

Woodall et al. (2013) esittävät laajemmin näiden kahden käsitteen eroavaisuuksia. He määrittelevät datan laadun arvioinnin prosessiksi laadun mittauksien saamiseksi ja datan laadun nykytilan määrittämiseksi. Yleensä datan laadun mittauksia toteutetaan määrittämällä arvoja eri mittareille, kuten laskemalla puuttuvat arvot tietokannasta. Mittauksia voidaan verrata viitearvoihin (esim. datan laatuvaatimuksiin), minkä avulla voidaan määrittää, kuinka montaa puuttuvaa arvoa voidaan sietää, jotta data sopii vielä käyttötarkoitukseen. Yleisen määritelmän mukaan datan laadun arvioinnin tarkoituksena on arvioida datan laadun mittauksia vaadittavien datan laatuparannuksien määrittämiseksi, vaikkakin

tarkkaa terminologiaa ei käytetä aina yhtenäisesti. Mittaamista voidaan pitää siis prosessina arvojen saamiseksi datan laadun ulottuvuuksille ja arvioinnissa verrataan näitä arvoja vertailuarvoihin laadun diagnosoinnin mahdollistamiseksi.

Esitettyjen näkemyksien myötä tässä työssä käytetään määritelmää, jonka mukaan mittauksista saadaan arvoja, joita arvioinnissa tarkastellaan tarvittavien datan laadun kehittämistoimenpiteiden määrittämiseksi. Tämän vuoksi tässä työssä käsitellään molempia datan laadun mittaamisen ja arvioinnin perusperiaatteita sekä keskeisimpiä menetelmiä. Datan laadun arviointiin keskitytään tarkemmin luvussa 4.

3.1 Mittaustyypit

Datan laatua voidaan mitata esimerkiksi datamallien, datan arvojen, datan alueiden, datan esityksen ja datan toimintaperiaatteiden näkökulmista. Datamallien laadun kohdalla voidaan mitata esimerkiksi joustavuutta, niiden kykyä heijastaa käyttäjien uusia vaatimuksia. Datan arvojen laatua voidaan mitata eri ulottuvuuksien avulla, kuten tarkkuuden ja täydellisyyden näkökulmista. Datan alueiden laadun mittaamisessa voidaan esimerkiksi tarkastella, miten hyvin yrityksen eri toimijat tekevät yhteistyötä yrityksen kattavien datan alatyyppien kanssa. Datan esityksen laadun mittaaminen edellyttää usein datan käyttäjien kanssa käytävää dialogia, koska esityksen laatuun vaikuttaa, miten käyttäjät omaksuvat sen. Datan toimintaperiaatteiden laadun mittaamista voidaan tarkastella esimerkiksi metadatan hallinnan, tietosuojan ja turvallisuuden näkökulmista. (Loshin 2001 ss. 210-227) Tässä työssä keskitytään datan arvojen laadun mittaamiseen.

Datan laadun mittaamiseen on käytännössä kaksi vaihtoehtoa, jotka ovat reaalimaailman testi ja arviointi. Reaalimaailman testin tapauksessa vahvistetaan, että vastaavatko datan esitykset todellisuutta vai ei. Reaalimaailman testi voidaan toteuttaa hyödyntämällä referenssidataa tai asiantuntijaryhmää, mutta asiantuntijoiden mielipide-erot saattavat kuitenkin johtaa epävarmuuteen. Reaalimaailman testi ei välttämättä ole toivottavaa esimerkiksi sen korkean hinnan vuoksi tai se ei ole mahdollista, jos ei ole pääsyä reaalimaailman arvoihin. Tämän vuoksi laadun arviointia sovelletaan useimmissa tapauksissa. Datan laadun arviointiin on olemassa kaksi järkevää tapaa, joista ensimmäinen on erilaisten sääntöjen tai rajoitteiden hyödyntäminen. Mittaukset ovat siis näiden erilaisten sääntöjen vahvistuksia. Tätä menetelmää hyödynnetään etenkin täydellisyyden ja johdonmukaisuuden mittaamiseen. Toinen tapa on mallin hyödyntäminen, missä kuvataan todellisuuden epävarmuutta ja laaditaan arvio tästä mallista. Kyseisessä menetelmässä mittaaminen edellyttää varmuuden mittaamista, että data todella on laadukasta. Tässä tapauksessa mittaukset ovat riippuvaisia käytetystä epävarmuusteoriasta. Esimerkiksi todennäköisyysmallin käyttäminen tarkoittaa sitä, että vastaava mittaus on määrällinen. (Bronselaer et al. 2018a) Toisaalta mittauksia voidaan jakaa myös staattisiin ja dynaamisiin mittauksiin. Staattisessa mittauksessa mitataan tutkittavan datan tilannekuvaa, kun taas dynaamisessa mittaamisessa mitataan dataa sen virran tiettyjen kohtien aikana. (Loshin 2001 s. 204)

Laadun eri ulottuvuuksilla on tyypillisesti erityiset luonteet, minkä vuoksi datan laadun mittaukset ovat hyvin hajanaisia ja heterogeenisiä. Tämän seurauksena yhteinen käsitys datan laadun mittaamisesta puuttuu. Esimerkiksi osa mittaamisen määritelmistä perustuu mittareihin, osa hyötylaskentaan ja toiset soveltavat datan toimintoja. Eri lähestymistapojen vertaaminen on vaikeaa, koska ne ilmaisevat laatua eri tavalla. (Bronselaer et al. 2018a)

Data on aineetonta, mutta sitä luodaan ja tallennetaan asiayhteydessä, mikä mahdollistaa sen mittaamisen. Esimerkiksi dataa voidaan määritellä, sääntöjä voidaan luoda kenttien täyttämiseen ja tietueita voidaan verrata toisiinsa. Mittaamiseen liittyy aina vertaileminen, sillä datan laadun mittaaminen vaatii sekä odotuksia datalle, että mittarin tarkkailemaan, missä määrin data vastaa näitä odotuksia. Datan ominaispiirteitä voidaan yhdistää datan kuluttajien odotusten tai muiden vaatimusten kanssa, ja tästä yhdistelmästä voidaan puolestaan luoda näihin ominaispiirteisiin tarkasti määriteltyjä mittauksia. (Sebastian-Coleman 2013 ss. 42-53)

Datan laadun mittaaminen voi olla objektiivista tai subjektiivista (Pipino et al. 2002; Batini et al. 2009; Sebastian-Coleman 2013 s. 60; Bronselaer et al. 2018a). Mittaaminen on objektiivista, kun se perustuu määrällisiin mittareihin (Batini et al. 2009). Objektiiviset mittarit mittaavat tehtävästä riippumattomia ominaispiirteitä. Kyseisiä mittareita voidaan käyttää ilman asiayhteystietoa datan käyttämisestä. Objektiiviseen datan laadun mittaamiseen kuuluu vähintään toinen kahdesta perusvertailusta: dataa voidaan mitata vertaamalla sitä selkeästi määriteltyyn standardiin tai itseensä ajan suhteen. Monimutkaisemmissa mittauksissa voidaan yhdistää nämä molemmat tyypit. Yksinkertainen esimerkki objektiivisesta mittauksesta on validien postinumeroiden tarkastelu. Jos ne on määritelty esimerkiksi viiden numeron pituisiksi, niin ne arvot eivät ole valideja, jotka eivät täytä tätä kriteeriä. (Sebastian-Coleman 2013 s. 60-61)

Pipino et al. (2002) puolestaan esittävät, että objektiivisia mittareita voidaan jakaa tehtävästä riippumattomien mittareiden lisäksi myös tehtävästä riippuvaisiin mittareihin. Tehtävästä riippumattomat mittarit kuvaavat datan tilaa ilman asiayhteysymmärrystä sovelluskohteesta, ja niitä voidaan hyödyntää mihin tahansa datajoukkoon, riippumatta kyseessä olevasta tehtävästä. Tehtävästä riippuvaiset mittarit, jotka sisältävät esimerkiksi organisaation liiketoimintasäännöt sekä yritys- ja hallintomääräykset, kehitetään tietyissä sovelluskonteksteissa.

Mittaaminen on subjektiivista puolestaan silloin, kun se perustuu datan käyttäjien ja hallinnoijien laadullisiin arviointeihin (Batini et al. 2009). Subjektiivisten ulottuvuuksien (esim. uskottavuus ja asiaankuuluvuus) mittaaminen edellyttää tietoa datan kuluttajilta, mitä voidaan kerätä esimerkiksi kyselyiden avulla. Subjektiiviset datan mittaukset heijastavat datan kuluttajien tarpeita ja kokemuksia. (Sebastian-Coleman 2013 s. 60) Molempien laadullisten subjektiivisten kyselyarviointien ja määrällisten objektiivisten mittareiden tapauksessa mittauksen tuloksena on datan arvo (Aljumaili et al. 2016).

Mittauksia voidaan jakaa myös rakenne- ja sisältöpohjaisiin mittauksiin. Rakennepohjaiset mittausmenetelmät pohjautuvat datan fyysisiin ominaisuuksiin, ja niissä oletetaan absoluuttisen standardin olemassaolo. Kyseiset menetelmät ovat objektiivisia, koska ne jättävät huomioimatta datan käyttöyhteyden. Rakennepohjaiset mittausmenetelmät perustuvat usein datan objektiivisiin ominaisuuksiin, kuten lukumäärien suhteisiin, aikamittauksiin tai virheiden määrään. (Ballou & Pazer 2003; Even & Shankaranarayanan 2005; Even & Shankaranarayanan 2009; Watts et al. 2009; Aljumaili et al. 2016; Bronselaer et al. 2018a)

Sisältöpohjaiset mittausmenetelmät, joita kutsutaan myös asiayhteydellisiksi arvioinneiksi, puolestaan juontuvat datan sisällöstä. Tyypillisesti kyseiset mittaukset heijastavat laatuvirheiden vaikutusta tietyssä käyttöympäristössä. (Ballou & Pazer 2003; Even & Shankaranarayanan 2005; Even & Shankaranarayanan 2009; Watts et al. 2009; Aljumaili et al. 2016) Tietyissä tapauksissa samaa ulottuvuutta voidaan mitata sekä objektiivisesti, että asiayhteydellisesti (Even & Shankaranarayanan 2009). Asiayhteydelliset mittaukset riippuvat käsiteltävän tehtävän vaatimuksista ja ominaisuuksista sekä käyttäjän ominaisuuksista (Bronselaer et al. 2018a). Laajemmin subjektiivisiin sisältöpohjaisiin tai asiayhteydellisiin mittauksiin vaikuttavia tekijöitä ovat:

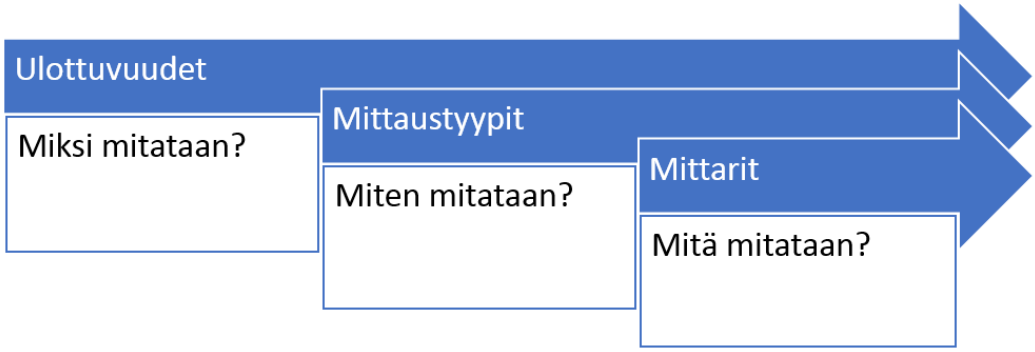
a) Laajuus: yksilöt, yksiköt ja koko organisaatio arvioivat datan laatua eri tavalla. Esimerkiksi yksittäinen käyttäjä voi olla kiinnostunut enemmän tietystä käytettävästä datasta, kun taas liiketoimintayksikkö voi tarkastella laatua tietovaraston näkökulmasta.
b) Tehtävä: tehtävän ominaisuudet vaikuttavat todennäköisesti laadun arviointiin. Esimerkiksi laatuvaatimukset poikkeavat strategisen päätöksenteon (esim. laajan liiketoiminta-alueen data) ja operatiivisten tarpeiden (esim. yksityiskohtainen data) välillä.
c) Rooli: eri sidosryhmät voivat korostaa laadun eri näkökohtia, riippuen heidän vastuustaan ja prosessivaiheesta, johon he osallistuvat.
d) Ajoitus: käyttäjät voivat arvioida laatua eri tavalla, kun käytön kiireellisyys on korkeampi.
e) Yksilö: asiayhteydelliseen mittaamiseen voivat vaikuttaa käyttäjän ominaisuudet, kuten motivaatio, osallistuminen ja kokemus. (Even & Shankaranarayanan 2005; Even & Shankaranarayanan 2007)

Datan laatu määritellään “sopivuudeksi käyttötarkoitukseen”, minkä näkökulmasta tekijät, kuten datan asiaankuuluvuus tehtävälle, käyttäjän kyky ymmärtää sitä ja tehtävän selkeys, vaikuttavat datan käytettävyyteen. Käytettävyyden näkökulmasta laadun arviointi on yleensä asiayhteydellistä, sillä data voi olla laadultaan hyväksyttävää yhdessä päätöksenteossa, mutta huonolaatuista toisessa päätöksenteossa. (Watts et al. 2009) Yleisesti on hyväksytty, että käytännön käyttötarkoituksen näkökulmasta datan ei tarvitse olla parasta laatua, jotta se olisi hyödyllistä. Tiettyyn sovelluskohteeseen liittyvä laatu on erilainen

kuin objektiivinen vaatimus, jonka mukaan datan on oltava niin hyvää kuin mahdollista. (Bronselaer et al. 2018a)

Taloudellisesta näkökulmasta objektiiviset mittaukset voidaan liittää kustannuksiin. Mitä enemmän datassa on virheitä, niin sitä enemmän aikaa ja vaivaa tarvitaan niiden korjaamiseen, mikä aiheuttaa enemmän kustannuksia. Toisaalta asiayhteydestä riippuen, datan laadun parantaminen vaikuttaa datan käytettävyyteen. Tämän vuoksi asiayhteydellinen mittaaminen voidaan liittää datan laadun parantamisesta saataviin hyötyihin. (Even & Shankaranarayanan 2009)

Ulottuvuudet kuvaavat datan laatuun liittyviä mitattavia näkökohtia ja tarjoavat perusteita mittauksien toteuttamiseen. Mittaustyypit puolestaan kuvailevat yleisiä tapoja ulottuvuuksien mittaamiseen. Yksityiskohtaiset mittarit taas kuvaavat, mitä dataa mitataan. (Sebastian-Coleman 2013 s. 48) Kuvassa 4 on esitetty ulottuvuuksien, mittaustyyppien ja mittarien suhdetta.



Kuva 4. *Ulottuvuuksien, mittaustyyppien ja mittarien suhde (mukaillen Sebastian-Coleman 2013 s. 48)*

Esimerkiksi validiuden mittaamisessa mittaustyyppinä voi olla datan arvojen vertaaminen valideihin arvoihin, kuten referenssitauluun tai matemaattiseen sääntöön. Yksityiskohtainen mittari voi olla tässä tapauksessa esimerkiksi myyntitulokoodien vertaaminen myyntitulokoodien referenssitauluun. (Sebastian-Coleman 2013 s. 48)

Jotkin ulottuvuuksista, kuten tarkkuus ja täydellisyys, vaativat objektiivisia mittauksia. Ne ovat mittauksia, jotka perustuvat dataan itsessään, huolimatta käytettävästä asiayhteydestä. (Watts et al. 2009) Kaikille ulottuvuuksille, kuten tulkittavuudelle, ei ole mahdollista määrittää objektiivista mittausta. Näissä tapauksissa laadun arviointi perustuu datan käyttäjien subjektiivisiin käsityksiin, ja tulokset voivat olla positiivisia tai negatiivisia riippuen käyttäjien tarpeista. (Cappiello et al. 2004) Lisäksi esimerkiksi asiaankuuluvuus ja uskottavuus vaihtelevat käytettävän asiayhteyden mukaan. Datan asiaankuuluvuus riippuu usein sovellettavasta tehtävästä, koska yhden tehtävän kannalta merkityksellinen data voi olla merkityksetöntä toiselle. Datan uskottavuutta on myös vaikea mitata objektiivi-

sesti, koska se riippuu usein käyttäjän kokemuksesta ja henkilökohtaisista mieltymyksistä. Esimerkiksi tietty data voi näyttää uskottavalta aloittelijalle, mutta se voi olla vähemmän uskottavaa asiantuntijalle. (Watts et al. 2009)

Objektiivisilla määrällisillä mittauksilla on kolme erilaista funktionaalista muotoa, jotka ovat yksinkertainen suhde, minimi- tai maksimiarvo sekä painotettu keskiarvo (Pipino et al. 2002; Cappiello et al. 2004; Even & Shankaranarayanan 2005). Yksinkertainen suhde mittaa vaadittujen tulosten suhdetta todellisiin tuloksiin. Se normalisoidaan yleensä välille 0-1, jossa 1 edustaa kaikkein toivotuinta tulosta ja 0 edustaa vähiten toivottua tulosta. (Cappiello et al. 2004; Even & Shankaranarayanan 2009; Aljumaili et al. 2016) Suhteet perustuvat laatuongelmien lukumäärän ja kokonaislukumäärän vertailuun (Even & Shankaranarayanan 2005).

Minimi- tai maksimiarvomuotoa käytetään niiden ulottuvuuksien kohdalla, mitkä edellyttävät useiden datan laatuindikaattorien yhdistämistä. Sitä käytetään myös datan laadun yhteenlasketun arvon esittämisessä yksittäisten ulottuvuuksien mukaan. (Cappiello et al. 2004)

Monimuuttujien tapauksessa käytetään myös painotettua keskiarvoa vaihtoehtona minimi- tai maksimiarvoille. Muuttujien painotettu keskiarvoa voidaan käyttää, jos yrityksellä on hyvä käsitys muuttujien merkityksestä ulottuvuuden kokonaisarvioinnissa. Normalisoidun tuloksen saamiseksi painoarvojen tulisi olla nollan ja yhden välillä. Yksinkertaista keskiarvoa voidaan käyttää, jos arvioidaan yksittäistä muuttujaa. Minimi- tai maksimiarvo ja painotettu keskiarvo ovat usein vaihtoehtoisia funktionaalisia muotoja. (Cappiello et al. 2004)

Mittaaminen on avaintoiminto datan laadunhallinnassa. Mittaamisen tarkoituksena on tyydyttää tiedon tarve tavoitteiden, riskien ja ongelmien hallitsemiseksi. (Aljumaili et al. 2016) Määrällinen mittaaminen on tunnistettu keskeiseksi tekijäksi onnistuneeseen datan laadunhallintaan (Even & Shankaranarayanan 2005). Se on kriittistä suurissa dataympäristöissä, sillä se voi auttaa luomaan realistisia laadun parantamistavoitteita, seuraamaan edistymistä, arvioimaan eri ratkaisujen vaikutuksia ja priorisoimaan parannustoimia (Even & Shankaranarayanan 2009; Aljumaili et al. 2016). Se on myös keskeistä datavirheiden ja niiden laajuuksien tunnistamisessa (Even & Shankaranarayanan 2009). Mittaukset tuovat huomiota alueille, joita johtohenkilöstö pitävät tärkeinä. Mittaukset tarjoavat tietoa esimerkiksi prosessien seuraamiseen ja projektien priorisoimiseen. Ennen ja jälkeen mittaukset tarjoavat tietoja laadun parantamiseen liittyvien toimintojen tehokkuudesta. Usein johtohenkilöstö harkitsee useita mahdollisia datan laadun kehittämisprojekteja, joten niillä on oltava hyvä tapa projektien priorisoimiseen. (Fisher et al. 2009)

3.2 Mittaamisen vaatimuksia

Mittauksien on oltava ymmärrettäviä, jotta ne voivat olla tehokkaita. Jos ihmiset eivät voi ymmärtää mitattavaa ominaisuutta, niin mittaukset eivät auta vähentämään epävarmuutta tai ne eivät ole edes hyödyllisiä, vaikka tarkastelun kohde on erittäin tärkeä. Mittaus on viestintäväline, ja samalla myös analysointityökalu. Datan kuluttajien täytyy ymmärtää mitattavan datan lisäksi, mitä mittaukset edustavat ja niillä on oltava tarpeeksi asiayhteyttä mittausten tulkitsemiseksi. Mittausten on oltava myös toistettavissa. Tärkein syy keskittyä mittaustyökaluihin ja mittausolosuhteisiin on johdonmukaisten mittaustulosten tuottaminen ja kaikkien tekijöiden ymmärtäminen, mitkä saattavat lisätä vaihtelevuutta mittaukseen. Jos mittaustyökalujen johdonmukaisuuteen ei voi luottaa, niin mittauksiin ei voi luottaa tai niillä on vain vähän merkitystä. (Sebastian-Coleman 2013 ss. 44-45) Laajemmin sanottuna mittaaminen vaatii ymmärrystä seuraavista asioista:

- Datan edustamat liiketoimintakäsitteet.
- Liiketoiminnalliset ja tekniset prosessit, jotka luovat dataa.
- Liiketoiminnalliset ja tekniset prosessit, jotka ylläpitävät, päivittävät tai poistavat dataa.
- Datamalli kohdejärjestelmässä, jossa dataa mitataan.
- Datan prosessointisäännöt kohdejärjestelmässä. (Sebastian-Coleman 2013 ss. 64-65)

Näistä kaikista viidestä aiheesta tarvitaan tietoa perustavanlaatuisen mittaamisen toteuttamiseksi. Liiketoimintakäsitteiden tuntemus on mittaamisen perustana, sillä se sisältää itse käsitteiden lisäksi myös sen, miten käsitteet ovat esitettyinä ja niiden suhteet. Dataa luodaan liiketoiminnallisten ja teknisten prosessien avulla, joten tietämys näistä prosesseista on välttämätöntä dataan liittyvien riskien ja lopulta datan laadun kehittämistapojen tunnistamiseksi. Data ei aina pysy alkuperäisessä muodossaan, sillä sitä voidaan muuttaa esimerkiksi päivitysten tai poistamisen myötä. Datamalli ja siihen liittyvät metadatat ovat kriittisiä työkaluja ymmärtämään, miten datan sisältö on tallennettu järjestelmään sekä tunnistamaan ja tarkentamaan dataan liittyviä odotuksia. On myös tunnettava datan prosessoimisessa käytettävät säännöt, jotta voidaan ymmärtää riskikohdat ja mittausmahdollisuudet. (Sebastian-Coleman 2013 s. 65)

Tarve varmistaa mittauksien ymmärrettävyys ja tulkittavuus on yksi muistutus metadatan kriittisyydestä (Sebastian-Coleman 2013 s. 44). Metadataa tarvitaan datan mittaamiseen, mutta metadatan hallinta on kuitenkin haaste useimmille organisaatioille. Monilla organisaatioilla ei ole esimerkiksi dokumentoitua datamallia. Ilman dokumentaatiota datan rakennetta on vaikea havaita ja se vaatii huomattavan määrän analyysejä. Jos metadataa ei ole olemassa datan laadun mittaamisen tukemiseksi, niin se on luotava osana mittausprosessia. (Sebastian-Coleman 2013 s. 65)

Kirjallisuuden monien lähestymistapojen mukaan datan laatua on mitattava mahdollisimman alhaisella tasolla ja sitten yhdistettävä korkeammille tasoille. Voi olla kuitenkin järkevää myös harkita heti korkeamman tason mittauksia. Esimerkiksi osoitetietojen täydellisyyden kohdalla usean attribuutin huomioiminen saattaa olla tarpeellista. Lisäksi useita attribuutteja kannattaa ottaa huomioon myös riippuvuussuhteiden tarkastelussa. Mittauksien laajuus on hyvä harkita etukäteen, koska sillä on seurauksia esimerkiksi mittausten koosteiden laskemisessa. (Bronselaer et al. 2018a)

Mittauksia voidaan toteuttaa datavirran eri vaiheissa. Käytännössä on kuitenkin hyvä huomioida muutamia perusperiaatteita mittauksien sijoittamiseen:

- Datan laatua tulisi mitata datavirran varhaisessa vaiheessa, jotta varmistetaan sen ja liiketoiminnallisten odotusten vastaavuus ennen datan prosessoimista.
- Jos dataa ei voida mitata ihanteellisessa kohdassa, niin sitä voidaan mitata myös lataustiedostosta. Se yksinkertaistaa datan mittaamista tapauksissa, joissa yhden tiedoston dataa jaetaan useisiin kenttiin.
- Alhaisemman riskin dataa (dataa, jota ei muuteta prosessoinnin aikana) voidaan mitata kerran ja seurata yksinkertaisemmin sen virtaamista. Seuranta varmistaa, että sen laatu ei heikkene. On kuitenkin huomioitava, että muut muutokset tietovirrassa voivat vaikuttaa myös dataan, jota ei ole tarkoitus muuttaa.
- Datan muunnokset ovat suurimpia riskejä datalle. Tämän vuoksi dataa on mitattava muunnosprosessien yhteydessä, jotta varmistetaan muutoksien tuovan odotettuja tuloksia. Dataa on mitattava mahdollisimman pian muunnoksien jälkeen. (Sebastian-Coleman 2013 ss. 79-80)

Datan hyödyllisyyden mittaamiseksi on usein tarpeellista mitata useita laadun ulottuvuuksia. Esimerkiksi osoitetietojen kohdalla tavoitteena voi olla mitata, onko osoitetiedot riittävän laadukkaita postin lähettämistä varten. Osoitetiedot voivat olla täydellisiä, mutta postitoimipaikan ja postinumeron väliset epäjohdonmukaisuudet voivat tehdä tiedoista hyödyttömiä kyseisen tavoitteen näkökulmasta. Nykyiset mittauksien lähestymistavat eivät ota huomioon, että erilaisilla mittauksilla yleensä on erilainen luonne ja tulkinta. (Bronselaer et al. 2018a)

Yleisesti mittaamiseen liittyvien vaatimusten lisäksi on myös esitetty tarkempia vaatimuksia itse mittareille. Niitä ovat mittarin minimi- ja maksimiarvojen olemassaolo, välimatka-asteikolliset arvot, konfigurointiparametrien ja mittarien arvojen määrittämisen laatu, mittarien arvojen yhdistäminen sekä mittarin taloudellinen tehokkuus. Ensimmäisen vaatimuksen mukaan mittauksen arvoja on rajoitettava alhaalta ja ylhäältä. Minimiarvot edustavat huonolaatuista dataa ja maksimiarvot puolestaan hyvää datan laatua. (Heinrich et al. 2018a)

Toinen vaatimus eli mittarin välimatka-asteikolliset arvot viittaavat merkityksellisten erojen ja aikavälien määrittämiseen. Ne ovat erittäin tärkeitä, kun arvioidaan, tulkitaan ja

vertaillaan erilaisten datan laadun kehittämismittauksien vaikutuksia taloudellisesti suuntautuneeseen datan laadunhallintaan. Esimerkiksi järjestysasteikollisessa mittarissa voi olla arvot ”erittäin hyvä”, ”hyvä”, ”kohtalainen”, ”huono” ja ”erittäin huono”. Tällöin ei ole kuitenkaan mahdollista yksilöidä arvojen erojen merkityksiä. Ei voida esimerkiksi määrittää, onko parannus ”erittäin huonosta” ”kohtalaiseen” yhtä paljon kuin parannus ”kohtalaisesta” ”erittäin hyväksi”. Välimatka-asteikollisessa mittaristossa puolestaan 0.2:n parannus on kaksi kertaa suurempi kuin 0.1:n parannus. (Heinrich et al. 2018a) Datan laadun ilmaiseminen välimatka-asteikossa [0,1] voi aiheuttaa ongelmia laadun mittausten tulkinnassa. Voi olla vaikea ymmärtää, mitä tarkoittaa esimerkiksi attribuutin tarkkuuden arvo on 0,7 tai mikä on välimatka-asteikon mittausyksikkö. Yleensä datan laadun osoittamiseen tarkoitetut numerot ovat vaikeasti tulkittavissa, eivätkä ne anna tietoa laadun heikkenemisen syistä tai mahdollisista toiminnoista datan laadun parantamiseksi. Tämän ongelman syynä voidaan pitää mittaamisen teoreettisen viitekehyksen puuttumista. Jos mittaamisen tarkoitusta ei ymmärretä, niin mittauksia ei voi helposti tulkita eikä niitä voi yhdistää. (Bronselaer et al. 2018a)

Kolmas vaatimus liittyy mittarin konfigurointiparametrien ja arvojen määrittämiseen objektiivisuuden, luotettavuuden ja pätevyyden mukaisesti. Objektiivisuus osoittaa, missä määrin parametrit, arvot ja niiden määritysmenetelmät ovat riippumattomia ulkoisista vaikutuksista. Objektiivisuutta rikotaan, jos liian harvat asiantuntijat antavat arvioita tai ulkoisia vaikutuksia ei minimoida. Subjektiivisten tulosten välttämiseksi ja puolueettomuuden varmistamiseksi datan laadun mittari ja sen konfigurointiparametrit ovat määriteltävä yksiselitteisesti objektiivisin menetelmin, esimerkiksi tilastollisten menetelmien avulla. Luotettavuus puolestaan käsitteellistää konfiguraatioparametrien tai metristen arvojen määrittämiseen käytettävien menetelmien tulosten toistettavuutta. Luotettavuutta voidaan analysoida eri mittausten tulosten korrelaation perusteella. Yleensä konfigurointiparametrien ja metristen arvojen luotettavuuden varmistamiseksi käytetään tietokantakyselyjä tai tilastomenetelmiä. Pätevyys taas viittaa siihen, että menetelmä mittaa sitä, mitä sen pitäisikin mitata. Tyypillisesti konfiguraatioparametrin tai mittarin arvon määrityksen pätevyyttä rikotaan, jos määritys on ristiriidassa tavoitteen kanssa. Yhteenvetoesimerkkinä voidaan todeta, että objektiivisuutta ja/tai luotettavuutta voidaan rikkoa konfiguraatioparametrien erilaisten asiantuntija-arvioiden vuoksi, ja pätevyyttä voidaan rikkoa epätäsmällisten määritelmien vuoksi. (Heinrich et al. 2018a)

Neljännen vaatimuksen mukaan datan laatumittarin on oltava soveltuva datan eri näkymätasoihin, sillä sitä on voitava hyödyntää yksittäisiin datan arvoihin sekä data-arvojen joukkoihin, kuten suhteisiin ja koko tietokantaan. Datan laatumittareita voidaan pitää matemaattisina funktiona, joiden pitää olla yhteensopivia eri datanäkymien tasojen koosteille, koska päätöstilanteet riippuvat yleensä suuren datajoukon laadusta. (Heinrich et al. 2018a) Ongelmia voi kuitenkin ilmentyä, kun yhdistetään mittauksia attribuuttien tasolta korkeammille tasoille. Pohjimmiltaan ne ovat erilaisia mittauksia ja niiden tulosten tulkinta voi olla myös erilainen, vaikka käytettäisiin samaa mitta-asteikkoa. Esimerkiksi

0,8:n täydellisyys voi olla hyväksyttävä tarkasteltavaan tehtävään nähden, kun taas 0,8:n tarkkuus voi tehdä datasta hyödytöntä. Perusongelma on se, että erilaiset mittaukset mittaavat olennaisesti jotain erilaista, minkä vuoksi eri mittausten tulos edellyttää erillistä käsittelyä. Vaikka mittaukset ilmaistaan samassa mitta-asteikossa, niin se ei kuitenkaan tarkoita sitä, että niitä voidaan tulkita samalla tavalla. Eri mittausten yhdistelmän tulisi olla tulkintatietoinen siinä mielessä, että yhdistelmätoiminto ottaa huomioon kunkin yksittäisen mittauksen tulkinnan. Yksittäiset yhdistettävät mittaukset tarjoavat perustan monimutkaisemmalle laadun mittaukselle. (Bronselaer et al. 2018a)

Viimeinen eli viides vaatimus on mittarin taloudellinen tehokkuus. Sen mukaan mittarista saatavien odotettujen tuottojen määrän on oltava suurempi kuin konfigurointiparametrien ja mittarien arvojen määrittämisen kustannuksien. Mittari voi kuitenkin olla arvokas teoreettisesta näkökulmasta, vaikka se ei täyttäisikään tätä vaatimusta. (Heinrich et al. 2018a)

3.3 Datan profilointi

Datan laadun kehittämiseen voi olla vaikea määrittää mittareita, koska ne sovelluskohtaisia. Yleinen menetelmä datan laadun määrittämiseksi on datan profilointi. (Andreescu et al. 2014) Datan profilointi on tietyntyyppinen data-analyysi, jota käytetään datajoukon ominaisuuksien etsimiseen ja luonnehtimiseen. Profilointi tarjoaa kuvan datan rakenteesta, sisällöstä, säännöistä ja suhteista tilastollisten menetelmien avulla. Tuloksena saadaan tietoa datan ominaisuuksista, kuten datan tyypeistä, kentän pituuksista, arvojoukoista, formaatti- ja sisältömalleista sekä epäsuorista säännöistä. (Sebastian-Coleman 2013 s. 49)

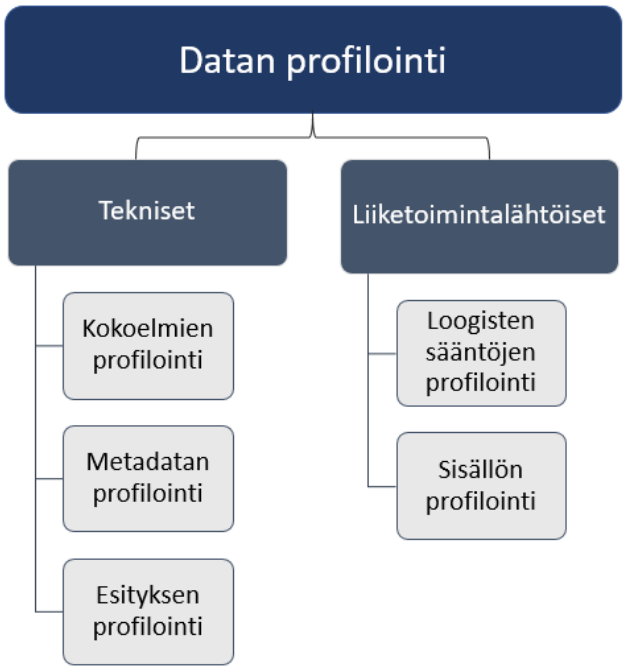
Datan profilointi on prosessi suurten datajoukkojen analysoimiseksi, mistä saadaan tilastollisia indikaattoreita datasta, kuten minimi-, maksimi- ja keskiarvo, keskihajonta sekä frekvenssi. Datan profiloimisen myötä saadaan myös metadataa, kuten datan tyyppi, pituus, erilliset arvot, ainutlaatuisuus, tyhjien arvojen esiintyvyys sekä tyypilliset merkkijonot. (Andreescu et al. 2014) Rakenteellisen, puolirakenteellisen ja rakenteettoman datan profiloinnin tuloksia voidaan hyödyntää datan hallinnassa, datan migraatioissa ja datan laadun valvonnassa (Dai et al. 2016).

Profiloinnin tuloksia voidaan verrata dokumentoituihin odotuksiin, tai ne voivat tarjota perustan ymmärryksen rakentamiseen datasta. Profilointi usein liitetään datan integraatioprojektien alkuvaiheeseen, jossa sitä käytetään datan tutkimiseen ja valmisteluun varastointia ja käyttämistä varten. Profilointia voidaan kuitenkin hyödyntää missä tahansa datan elinkaaren vaiheessa. Useimmat dataresurssit hyötyvät säännöllisestä uudelleen profiloimisesta laadun tason varmistamiseksi. Jos taas laatu on muuttunut, niin on huomioitava myös mahdolliset muutokset liiketoimintaprosesseissa. (Sebastian-Coleman 2013 s. 49)

Datan profiloinnin keskeisimmät menetelmät voidaan jakaa kolmeen ryhmään, jotka ovat rakenteen, sisällön ja suhteiden analysointi (Dorr & Murnane 2011; Andreescu et al. 2014; Mahanti 2014; Azeroua et al. 2018).

1. Rakenteen analysoinnissa tarkistetaan datan ja metadatan vastaavuus sekä datamallin pätevyys (Azeroua et al. 2018). Siinä tarkastellaan datan perusominaisuuksia ja arvioidaan niiden sopivuutta käyttötarkoitukseen. Esimerkiksi hahmonsovituksen (*engl. Pattern matching*) avulla voidaan selvittää, onko datan arvot määritetyssä muodossa. (Dorr & Murnane 2011) Tämän lisäksi yleisiä rakenteen analysoimisen tekniikoita ovat validointi metadatan avulla ja tilastotieteen hyödyntäminen (Mahanti 2014).
2. Sisällön analysoinnissa tarkistetaan datan täydellisyys, virheettömyys, ajantasaisuus ja erilaisten sääntöjen mukaisuus (Azeroua et al. 2018). Sisällön analysoinnissa tarkastellaan tarkemmin yksittäisiä dataelementtejä ja se voi esimerkiksi paljastaa arvojen samankaltaisuuksia ja eroja datalähteiden sisällä tai niiden välillä (Mahanti 2014). Sen tekniikat auttavat löytämään epästandardia dataa tai poikkeavuuksia (Dorr & Murnane 2011).
3. Suhteiden analysoinnissa tarkistetaan sarakkeiden ja taulukoiden keskeisimmät avainsuhteet (Azeroua et al. 2018). Sen ydin on ymmärtää, minkälaisia suhteita datalla on erilaisten sovellusten ja lähteiden välillä. Esimerkiksi asiakasnumero voi toimia asiakkaan ensisijaisena tunnisteena. Voi kuitenkin löytyä tapauksia, joissa myyntitilauksella ei ole asiakasnumeroa tietokannassa tai vaihtoehtoisesti tietokannasta voi löytyä duplikaatteja. (Mahanti 2014)

Datan profiloinnin tehtäviä voidaan jaotella myös viiteen ensisijaiseen tehtävään. Ensimmäinen on metadatan profilointi, jonka avulla saadaan tietoa datarakenteista, luojista, luomisajoista sekä pää- ja viiteavaimista. Toinen tehtävä on esityksen profilointi, jossa etsitään datamalleja, kuten teksti-, aika- ja numeromalleja. Kolmas on sisällön profilointi, jossa tarkistetaan datan perustiedot, kuten tarkkuus, ajantasaisuus ja tyhjät arvot. Neljäs tehtävä on datajoukkojen tai -kokoelmien profilointi, jossa analysoidaan dataryhmiä. Sen avulla saadaan tietoa ainutlaatuisuudesta, rivien määristä, minimi- ja maksimiarvoista sekä keskiarvoista. Viides tehtävä on loogisten sääntöjen profilointi, missä dataa tarkastellaan liiketoiminnan loogisten sääntöjen, liiketoimintasanastojen tai liiketoimintasääntöjen perusteella. (Dai et al. 2016) Kuvassa 5 on jaoteltu nämä dataprofiloinnin tehtävät teknisiin ja liiketoimintalähtöisiin.



Kuva 5. *Dataprofiloinnin tehtäviä (mukaillen Dai et al. 2016)*

Profilointitekniikat voidaan jakaa kahteen kategoriaan, manuaalisiin ja automatisoituihin. Manuaaliset tekniikat edellyttävät ihmisiä, jotka tarkastelevat datan tilaa kyselyiden avulla. Tämä lähestymistapa sopii pienille, yhdestä lähteestä tuleville ja suhteellisen yksinkertaisille datajoukoille, joissa on vähemmän kuin 50 datakenttää. Automaattisissa tekniikoissa puolestaan hyödynnetään ohjelmistotyökaluja yhteenvetotilastojen ja analyysien keräämiseen. Kyseiset työkalut sopivat parhaiten projekteihin, joissa on satoja tuhansia tietueita sekä useita kenttiä ja lähteitä. Kun datan profilointiprosessi on valmis ja kaikki ongelmat ovat tunnistettu, niin erityistä huomiota on kiinnitettävä datan siivoamiseen. Datan siivoamisen avulla voidaan poistaa virheitä ja epäjohdonmukaisuuksia, minkä myötä parannetaan datan laatua. (Andreescu et al. 2014)

3.4 Mittarit

Datan laadun ulottuvuudet tarjoavat tiettyjä näkökulmia datan laatuun. Kirjallisuudesta löytyy useita erilaisia mittareita näiden ulottuvuuksien määrälliseen mittaamiseen. (Heinrich et al. 2018a) Datan laadun mittaamisen menetelmissä on joustavuutta, sillä jokaista ulottuvuutta voidaan mitata useilla eri tavoilla (Aljumaili et al. 2016). Usein vaikein tehtävä mittaamisessa on yrityksen sovelluskohteeseen liittyvän ulottuvuuden tarkka määrittely. Tämän jälkeen mittarin muodostamista voidaan pitää suoraviivaisena. (Pipino et al. 2002)

Datan laadun mittareita tarvitaan kahdesta keskeisestä syystä, joista ensimmäinen on datapohjaisen päätöksenteon tukeminen. Siihen tarvitaan perusteltuja datan laadun mittareita, joiden avulla päätöksentekijät voivat arvioida datan luotettavuutta. Toinen keskei-

nen syy on mittareiden arvojen hyödyntäminen taloudellisesti suuntautuneen datan laadunhallinnan tukemiseen. Tässä asiayhteydessä datan laadun kehittämismittauksia tulisi hyödyntää vain, jos hyödyt ovat suuremmat kuin niihin liittyvät kustannukset. Tarvitaan perusteltuja datan laadun mittareita laadun tason määrittämiseksi, jotta voidaan analysoida, mitkä datan laadun kehittämistoimenpiteet ovat taloudellisesti tehokkaita. (Henrich et al. 2018)

Mittarit ovat tärkeä osa datan laadunhallintaa ja ne ovat usein liiketoiminnan korkeita prioriteetteja. Ne eivät ole datan laatuprojektien viimeinen vaihe, vaan niiden tarkoituksena on tuoda näkyvyyttä datan laatuongelmien estämiseksi. Mittarit ovat hyödyllisiä:

- Mielipiteiden korvaamisessa tosiasioilla.
- Resurssien kohdistamisessa.
- Ongelmien lähteiden tunnistamisessa.
- Ratkaisujen tehokkuuden vahvistamisessa.
- Liiketoimintatavoitteita tukevan käyttäytymisen kannustamisessa datan laadun välityksellä. (McGilvray 2008 s. 269)

Laadun seurantaa varten tulee valita muutamia keskeisimpiä mittareita, koska kaikkia mittareita ei voi eikä pitäisi seurata. Mittareiden valinnassa voidaan huomioida useita eri tekijöitä, kuten mittarin prioriteetti, mittausmenetelmä, mittaustiheys, kustannusten ja hyötyjen suhde sekä huomiotta jättämisen riski. (Umar et al. 1999)

Yleensä mittareiden arvot vaihtelevat välillä 0-1, jossa 1 tarkoittaa toivottua datan laatua ja 0 viittaa vähiten toivottuun laatuun (Pipino et al. 2002; Cappiello et al. 2004; Even & Shankaranarayanan 2009; Blake & Mangiameli 2011; Aljumaili et al. 2016). Tyypilliset datan laadun mittarit pohjautuvat kaavaan (1).

$$Suhde = 1 - \left[\frac{Epämieluisten\ tuloksien\ lukumäärä}{Kokonaislukumäärä}\right] \quad (1)$$

Tässä tapauksessa mittari on objektiivinen, sillä se koostuu vain kriteerin täyttävien datayksiköiden lukumäärän laskemisesta (Pipino et al. 2002; Caballero et al. 2007; Aljumaili et al. 2016). Mittaria voidaan hyödyntää myös pitkittäisvertailuissa, joka havainnollistaa jatkuvan parantamisen suuntauksia (Pipino et al. 2002).

Täydellisyys (*engl. Completeness*) voidaan määritellä datasta puuttuvien arvojen funktiona (Pipino et al. 2002). Datan arvolla ”NULL” voidaan esittää kaikkia niitä arvoja, jotka puuttuvat tai ovat tuntemattomia. Kaavassa (2) on esitetty täydellisyyden mittari.

$$Täydellisyys = 1 - \frac{T_R}{N_R} = \frac{N_R - T_R}{N_R} \quad (2)$$

Kaavassa (2) N_R tarkoittaa rivien kokonaismäärää ja T_R tarkoittaa ”NULL”-arvon sisältävien rivien määrää. (Heinrich et al. 2018a) Täydellisyyttä voidaan tarkastella myös tosi- ja epätosi- lausekkeilla, mikä on esitetty kaavassa (3).

$$Täydellisyys = \begin{cases} Epätosi, jos\ v_{col,row} = null\ tai\ \ v_{col,row} \in \{NaN, -, \dots\} \\ Tosi\ muuten \end{cases} \quad (3)$$

Sarake tai rivi ($v_{col,row}$) on epätäydellinen, jos siitä puuttuu arvoja tai ne on merkitty tyhjiksi. On hyvä huomioida, että puuttuvia arvoja voidaan määrittää useilla eri tavoilla. (Bors et al. 2018)

Validiutta (*engl. Validity*) voidaan mitata esimerkiksi tarkistamalla datan yhteensopivuus automaattisesti tai manuaalisesti määritettyjen datatyyppien kanssa (Bors et al. 2018). Kaavassa (4) on esitetty validiuden mittari.

$$Validius = \begin{cases} Tosi, jos\ tyyppi(v_{col,row}) = tyyppi,\ \ tyyppi \in \{numeerinen, merkkijono \dots\} \\ Epätosi\ muuten \end{cases} \quad (4)$$

Kaavan (4) tulokseksi saadaan epätosi, jos datan tyyppi ei kuulu ennalta määriteltyjen datatyyppien joukkoon. (Bors et al. 2018) Ajantasaisuus (*engl. Currency*) voidaan laskea toimitusajan, syöttöajan ja iän avulla. Toimitusaika kertoo, koska datan käyttäjä on vastaanottanut datan. Syöttöaika puolestaan viittaa aikaan, jolloin järjestelmä on vastaanottanut datan. Ikä viittaa datan ikään, kun järjestelmä vastaanotti sen ensimmäisen kerran. (Ballou et al. 1998; Pipino et al. 2002) Kaavassa (5) on esitetty ajantasaisuuden mittari.

$$Ajantasaisuus = (Toimitusaika - Syöttöaika) + Ikä \quad (5)$$

Oikea-aikaisuus (*engl. Timeliness*) voidaan laskea ajantasaisuuden ja volatiliteetin avulla. Volatiliteetti on attribuutin arvon maksimiaika, jolloin sitä voidaan pitää vielä ajantasaisena. (Ballou et al. 1998; Pipino et al. 2002) Kaavassa (6) on esitetty oikea-aikaisuuden mittari.

$$Oikea - aikaisuus = max\left[1 - \frac{Ajantasaisuus}{Volatiliteetti}, 0\right]^s \quad (6)$$

Eksponenttia voidaan hyödyntää herkkyyskertoimena. Sen arvo on tehtävästä riippuvainen ja se heijastaa analyytikon arviointia. (Ballou et al. 1998; Pipino et al. 2002)

Datan uskottavuus (*engl. Believability*) ja luotettavuus (*engl. Reliability*) tarkoittavat, missä määrin dataa pidetään todenmukaisena. Uskottavuutta voidaan tarkastella esimerkiksi arvioimalla datalähteen uskottavuutta, vertailemalla yleisesti hyväksyttyyn standardiin ja aikaisempiin kokemuksiin. Jokainen näistä muuttujista mitoitetaan asteikolla 0-1

ja yleiseksi uskottavuudeksi määritetään näiden kolmen minimiarvo. Vaihtoehtoisesti uskottavuus voidaan laskea myös näiden komponenttien painotettuna keskiarvona. (Pipino et al. 2002) Uskottavuudesta voidaan esittää myös mittari kaavan (7) muodossa.

$$Uskottavuus = \sum_{i=1}^{n} Q_i \tag{7}$$

Kaavassa (7) i tarkoittaa kysymystä koskien datajoukon uskottavuutta ja n viittaa kysymysten määrään. (Heinrich et al. 2018a)

Oikeellisuutta (*engl. Correctness* tai *Free of error*) voidaan mitata jakamalla virheellisten datayksiköiden lukumäärä datayksiköiden kokonaismäärällä ja vähentämällä se yhdestä. Tähän vaaditaan selkeitä kriteerejä virheiden määrittämiseen. (Pipino et al. 2002) Toisaalta oikeellisuuden mittaamisessa voidaan myös tarkastella, miten hyvin data vastaa reaalimaailman arvoja. Oikeellisuuden mittari on esitetty kaavassa (8).

$$Oikeellisuus = \frac{1}{d(\omega, \omega_m) + 1} \tag{8}$$

Kaavassa (8) ω tarkoittaa arvioitavan datan arvoa, ω_m tarkoittaa vastaavaa reaalimaailman arvoa ja d tarkoittaa etäisyysmittaa, kuten euklidista etäisyyttä. (Heinrich et al. 2018a)

Johdonmukaisuutta (*engl. Consistency*) voidaan tarkastella useista näkökulmista, joista yksi on samojen data-arvojen johdonmukaisuus eri tietokantatauluissa. Kuten aiemmin on mainittu, myös tätä voidaan mitata laskemalla tiettyjen johdonmukaisuusrikkomusten suhde johdonmukaisuustarkistuksien kokonaismäärään ja vähentämällä se yhdestä. (Pipino et al. 2002) Johdonmukaisuutta voidaan tarkastella myös assosiaatiosääntöjen johdonmukaisuuden perusteella, minkä mukainen mittari on esitetty kaavassa (9).

$$Johdonmukaisuus(t) = \sum_{r \in R} \begin{cases} w^+(r), jos\ t\ täyttää\ r \\ w^-(r), jos\ t\ rikkoo\ r \\ w^0(r), jos\ r\ ei\ päde \end{cases} \tag{9}$$

Kaavassa (9) R tarkoittaa assosiaatiosääntöjen joukkoa, $w^+(r)$ ja $w^-(r)$ kuvaa täytetyn ja rikotun assosiaatiosäännön pisteytyksen ja $w^0(r)$ on pisteytys soveltumattomalle assosiaatiosäännölle. (Heinrich et al. 2018a)

Datan tarkoituksenmukainen määrä (*engl. Appropriate amount of data*) tarkoittaa, että dataa ei ole liian vähän eikä liian paljon. Yleinen mittari tämän laskemiseen minimiarvo seuraavista suhteista: suhde toimitettujen datayksiköiden ja tarvittavien datayksiköiden lukumäärän välillä sekä suhde tarvittavien datayksiköiden ja toimitettujen datayksiköiden lukumäärän välillä. (Pipino et al. 2002) Taulukossa 6 on esitetty erilaisia mittareita ulottuvuuksille, mitä ei vielä aiemmin olla tuotu esille.

Taulukko 6. *Ulottuvuuksia ja niiden mittareita (mukaillen Batini et al. 2009)*

Ulottuvuus	**Mittari**
Ajantasaisuus	(1) Aika, jolloin data tallennetaan järjestelmään – aika, jolloin dataa päivitetään reaalimaailmassa (2) Aika viimeisimmästä päivityksestä (3) Pyyntöaika – päivitysaika (4) Käyttäjäkysely
Asiaankuuluvuus	Käyttäjäkysely
Johdonmukaisuus	(1) Rajoitteita rikkovien arvojen määrä (2) Käyttäjäkysely
Maine	Käyttäjäkysely
Objektiivisuus	Käyttäjäkysely
Oikea-aikaisuus	(1) Prosenttiosuus prosessitoteutuksista, jotka pystytään suorittamaan vaaditussa ajassa (2) Käyttäjäkysely
Saatavuus	(1) Max(0; 1-(toimitusaika – pyyntöaika)/(määräaika – pyyntöaika) (2) Käyttäjäkysely
Tarkkuus	(1) Syntaktinen tarkkuus: tietokantaan tallennettujen arvojen ja oikeiden arvojen välinen etäisyys (2) Toimitettujen tarkkojen arvojen määrä (3) Käyttäjäkysely
Tulkittavuus	(1) Tulkittavien datojen määrä (2) Keskeisten arvojen dokumentaatio (3) Käyttäjäkysely
Turvallisuus	(1) Heikkojen sisäänkirjautumisten määrä (2) Käyttäjäkysely
Täydellisyys	Käyttäjäkysely

Yritykset haluavat käytännössä kehittää yhden koostetun mittarin datan laadusta, datan laadun indeksin. Yksiarvoinen koostemittari on kuitenkin altis kaikille niille puutteille, jotka liittyvät laajasti käytettyihin indekseihin, kuten kuluttajahintaindeksiin. Monet muuttujat olisivat tällöin subjektiivisia, ja vaikeuksia toisivat myös eri asteikkotyyppien (järjestysluku, aikaväli, suhde) yhdistämiset. Jos oletukset ja rajoitukset ymmärretään ja indeksiä tulkitaan niiden mukaisesti, niin se voi auttaa yrityksiä arvioimaan niiden datan laadun tilaa. Kyseinen indeksi voi myös helpottaa datan laadun tilan viestimistä ylimmälle johdolle ja tarjota vertailevia arviointeja ajan suhteen. (Pipino et al. 2002)

3.5 Mittareiden esittäminen

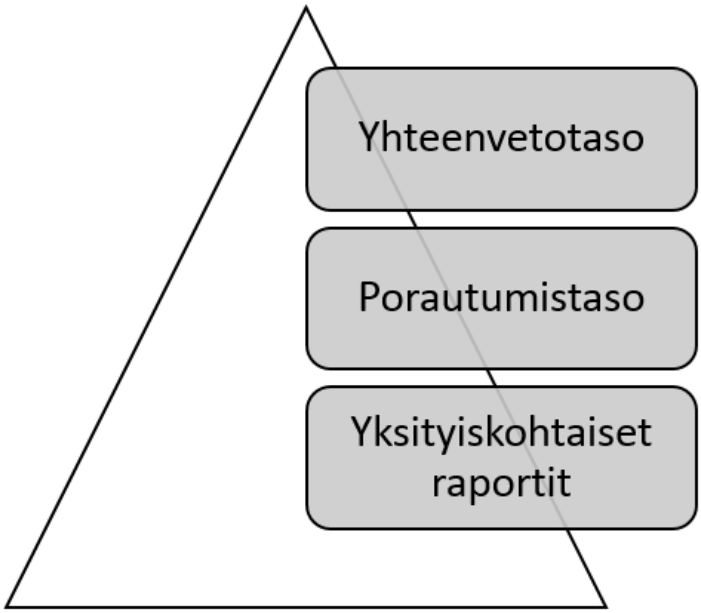
Käyttäjälle tuntemattoman datajoukon mittaaminen tai usein päivittyvän datan laadun ja rakenteen muutoksien tunnistaminen voi olla vaikea tehtävä. Yksi lähestymistapa datan laadun mittaamiseen on yhteenvetovisualisointien hyödyntäminen datan jakautumisen ja poikkeavuuksien tunnistamiseen. Yhteenvetovisualisoinneilla ei kuitenkaan ole joustavuutta datan laatua koskevien erilaisten näkökohtien korostamiseen. Automaattisesti lasketut laatumittarit voivat helpottaa laadun mittaamista ja nopeuttaa validointia. Yksittäisiin laatutarkastuksiin verrattuna laatumittareita voidaan käyttää datan erilaisten ominaispiirteiden samanaikaiseen validointiin. Yleistetyt mittaukset eivät kuitenkaan usein riitä määrittelemään tiettyyn sovelluskohteeseen liittyviä laatuongelmia. Asiayhteyteen liittyvien datan ominaispiirteiden mittaaminen edellyttää mittareiden sopeuttamista ja kustomointia. (Bors et al. 2018)

Datan esikäsittelyn aikana analyytikot käyttävät huomattavan määrän aikaa ja vaivaa datan laadun profiloimiseen sekä datan siivoamiseen ja muokkaamiseen analyysejä varten. Vaikka datan laadun mittarit tukevat datajoukon laadun mittaamista ja arviointia, niin hyvin harvat lähestymistavat ottavat huomioon analyytikkojen visuaalisen tukemisen mittareiden muokkaamisessa ja soveltamisessa. Visuaaliset lähestymistavat voivat kuitenkin helpottaa käyttäjien osallistumista datan laadun mittaamiseen. Datan laadun mittaamisen työkalulle tai ympäristölle voidaan määrittää viisi vaatimusta:

1. Mukautettavat laatumittarit. Datan laatumittareiden pitäisi asianmukaisesti heijastaa käsiteltävän datan laatua. Tämän saavuttamiseksi käyttäjien on kyettävä mukauttamaan laatumittareita tiettyjä datajoukkoja varten. Ennalta määriteltyjen käyttövalmiiden mittareiden parametrien pitäisi olla helposti säädettävissä, jotta käytön joustavuus säilyisi.
2. Datan laadun yleiskuvaus. Datan laadusta tulisi esittää visuaalinen yleiskuvaus. Sen tulisi erityisesti välittää tietoa mahdollisista virheistä, joita on havaittu datajoukossa.
3. Virheilmoitukset. Yksityiskohtaiset tiedot mahdollisesta huonosta datasta tulisi viestittää käyttäjille. Tämän tiedon tulisi helpottaa virhelähteiden tunnistamista.
4. Virhejakauma. Virheet esiintyvät harvoin yksittäin datassa. Tämän vuoksi käyttäjien pitäisi pystyä tarkastelemaan virheiden jakautumista, mikä saattaa paljastaa datasta kaavoja. Lisäksi työkalun pitäisi helpottaa virheiden korrelaatioiden havaitsemista useista datataulukon sarakkeista.
5. Datan tutkiminen. Käyttäjän tulisi pystyä siirtymään laatuongelmia sisältäviin arvoihin, jotta huonolaatuisen datan tarkastaminen helpottuisi. (Bors et al. 2018)

Mittareiden esittämisessä voidaan ottaa huomioon myös eri yleisöjen tarpeet. Esimerkiksi yhteenvetotasossa voidaan näyttää, kuinka hyvin keskeiset attribuutit noudattavat yrityksen standardeja. Porautumistaso voi sisältää keskeisimpien attribuuttien datan laadun statuksen eri järjestelmien mukaan. Jos mittareiden raportteja esitetään verkkosivustolla,

niin porautumistaso voi sisältää linkkejä yksityiskohtaisiin raportteihin, jotka sisältävät todelliset poikkeustilastot, laatuongelmat ja juurisyiden analysoinnin statuksen. (McGilvray 2008 s. 270) Kuvassa 6 on esitetty nämä kolme mittareiden esittämisen tasoa.



Kuva 6. *Mittareiden esittämisen kolme tasoa (mukaillen McGilvray 2008 s. 270)*

Yhteenvetotason mittarit tarjoavat helposti tulkittavan visuaalisen näkymän mittareista, kuten tavoitteista ja todellisesta datan laadun statuksesta. Status ilmaisee mittareiden tilanteen helposti ymmärrettävien termien avulla. Esimerkiksi vihreää väriä voidaan käyttää ”tuloksiin, jotka täyttävät tai ylittävät tavoitteen”, keltaista ”epäonnistuneisiin tuloksiin” ja punaista ”tuloksiin, jotka ovat sietorajojen ulkopuolella”. Yhteenvetoja voidaan näyttää tietyn ajankohdan mukaan tai niihin voidaan sisällyttää myös trendejä ja historiaa. (McGilvray 2008 s. 270)

Porautumistaso (*engl. Drilldown*) on keskitason näkymä, joka tarjoaa lisätietoja yhteenvetotason mittareista. Yhteenvetotasossa voi olla esimerkiksi kohta ”näytä lisätietoja mittareista”, mikä ohjaa porautumistasolle. (McGilvray 2008 s. 270)

Mittarit perustuvat yksityiskohtaisiin mittauksiin, jotka esitetään yksityiskohtaisissa raporteissa. Johtotaso ei yleensä katso yksityiskohtaisia raportteja, mutta niiden tulisi kuitenkin olla saatavilla, jos kysymyksiä herää mittareiden tarkkuudesta. Projektiryhmän tulisi käyttää yksityiskohtaisia raportteja poikkeamien seuraamiseen ja korjaamiseen. (McGilvray 2008 s. 270)

Tutkijat ovat myös ehdottaneet, että objektiivisia laadun mittauksia liitetään päätöksenteossa käytettävään dataan, jotta päätöksentekijät saisivat lisäinformaatiota. Näitä mittauksia voidaan kutsua esimerkiksi datamerkinnöiksi, datan laadun informaatioksi ja laadun metadataksi. (Watts et al. 2009)

4. DATAN LAADUN ARVIOINTI

Datan laadun arviointi- termillä voidaan viitata joukkoon prosesseja, jotka ovat suunnattu datan tilan ja arvon arvioimiseen organisaation sisällä. Datan laadun arvioinnin tarkoituksena on tunnistaa datavirheitä ja mitata erilaisten datapohjaisten liiketoimintaprosessien vaikutuksia. Molemmat virheiden tunnistaminen ja niiden vaikutusten ymmärtäminen ovat kriittisiä. Datan laadun arviointi voidaan toteuttaa eri tavoin yksinkertaisesta laadullisesta arvioinnista yksityiskohtaiseen määrälliseen mittaukseen. Arviointeja voidaan tehdä yleisen tietämyksen, ohjaavien periaatteiden tai tiettyjen standardien perusteella. Dataa voidaan arvioida yleisen sisällön makrotasolla tai tiettyjen kenttien tai arvojen mikrotasolla. Datan laadun arvioinnin keskeisenä tarkoituksena on ymmärtää datan tilaa suhteessa odotuksiin tai tiettyihin tarkoituksiin tai molempiin, ja tehdä päätelmiä, täyttääkö se odotuksia tai tietyn käyttötarkoituksen vaatimuksia. Tähän prosessiin kuuluu aina myös tarve ymmärtää, miten tehokkaasti data esittää kohteita, tapahtumia ja käsitteitä, joita se on suunniteltu esittämään. (Sebastian-Coleman 2013 ss. 46-47)

Tässä työssä datan laadun arvioinnilla tarkoitetaan prosessia, jossa hyödynnetään datan laadun mittauksia laadun diagnosoimiseksi ja tarvittavien datan laadun kehittämistoimenpiteiden määrittämiseksi. Tässä luvussa esitetään arviointimenetelmien vertailuperiaatteiden lisäksi neljä erilaista arviointimenetelmää, joista viimeisenä esitettävää Hybridi-menetelmää hyödynnetään empiriassa sen kokonaisvaltaisuuden ja joustavuuden vuoksi.

4.1 Arviointimenetelmien vertailuperiaatteet

Datan laadun kehittämissuunnitelma on aloitettava skenaarioiden arvioinnilla laatuongelmien juurisyiden tunnistamiseksi. Arvioinnin suorittamiseen tarvitaan arvoja datan laadun mittauksista. Näiden mittauksien keskeisimpänä tarkoituksena on määrällisen merkityksen tarjoaminen siitä, kuinka monessa laadun ulottuvuudessa päästään tavoitteeseen. (Caballero et al. 2007)

Datan laadun menetelmä voidaan määritellä joukoksi ohjeita ja tekniikoita, jotka määrittelevät prosessin datan laadun arvioimiseksi ja parantamiseksi (Batini et al. 2009). Tässä työssä kyseisillä menetelmillä tarkoitetaan datan laadun arviointimenetelmiä. Datan laadun arviointimenetelmien analysoimiseen ja vertailemiseen on olemassa useita näkökulmia, joita ovat:

1. Menetelmään kuuluvat vaiheet.
2. Strategiat ja tekniikat, joita käytetään datan laadun arviointiin ja parantamiseen.
3. Ulottuvuudet ja mittarit, jotka ovat valittu datan laadun tason arviointiin.
4. Kustannustyypit, jotka liittyvät datan laatuongelmiin.
5. Datan tyypit, jotka otetaan menetelmissä huomioon.

6. Tietojärjestelmien tyypit, jotka käyttävät, muokkaavat ja hallinnoivat dataa.
7. Organisaatiot, jotka osallistuvat datan luomiseen tai päivittämiseen liittyviin prosesseihin.
8. Prosessit, jotka luovat tai päivittävät dataa.
9. Palvelut, joita prosessit tuottavat. (Batini et al. 2009)

Menetelmät eroavat toisistaan siinä, miten ne huomioivat nämä näkökulmat. Yleisimmissä tapauksissa datan laadun menetelmän toiminnat koostuvat kolmesta vaiheesta. Ensimmäinen on tilan rekonstruktio, minkä tarkoituksena on kerätä kontekstuaalista tietoa organisaation prosesseista ja palveluista, datan keräämisestä ja siihen liittyvistä hallintamenettelyistä sekä laatuongelmista ja niiden kustannuksista. Tämä vaihe voidaan ohittaa, jos kontekstuaalista tietoa on saatavilla aiemmista analyyseistä. Toinen vaihe on mittaus, jossa mitataan datan laatua asianmukaisten ulottuvuuksien kanssa. Kolmas vaihe on kehittämistoimenpiteet, jossa valitaan vaiheet, strategiat ja tekniikat datan laatutavoitteiden saavuttamiseksi. (Batini et al. 2009) Näiden kolmen yleisimmän vaiheen sisältöä voidaan jakaa myös pienempiin osiin. Mittaus-vaiheeseen kuuluvat:

- Datan analysointi, missä tutkitaan datan malleja ja suoritetaan haastatteluja ymmärryksen muodostamiseen datasta sekä siihen liittyvistä arkkitehtuuri- ja hallintasäännöistä.
- Datan laadun vaatimusten analysointi, missä kartoitetaan datan käyttäjien ja hallinnoijien mielipidettä laatuongelmien tunnistamiseksi ja uusien laatutavoitteiden asettamiseksi.
- Kriittisten alueiden tunnistaminen, missä valitaan tärkeimmät tietokannat ja datavirrat määrälliseen arviointiin.
- Prosessien mallintaminen, mikä tarjoaa mallin prosesseista, jotka tuottavat tai päivittävät dataa.
- Laadun mittaaminen, missä valitaan datan laadun vaatimusten analysointi –vaiheessa tunnistetut laatuongelmiin liittyvät ulottuvuudet ja määritellään niitä vastaavat mittarit. Mittaus voi olla objektiivista, kun se perustuu määrällisiin mittareihin tai subjektiivista, kun se perustuu datan käyttäjien laadullisiin määrityksiin. (Batini et al. 2009)

Kehittämistoimenpiteisiin puolestaan kuuluvat:

- Kustannusten arviointi, missä arvioidaan suoria ja epäsuoria datan laadun kustannuksia.
- Prosessin vastuiden määrittely, mihin kuuluu prosessin omistajien tunnistaminen ja heidän vastuualueiden määrittäminen datan tuottamisessa ja hallitsemisessa.
- Datan vastuiden määrittely, mihin kuuluu datan omistajien tunnistaminen ja heidän datan hallinnan vastuualueiden määrittäminen.

- Datan laatuongelmien syiden tunnistaminen, missä valitaan datan kehittämisen strategiat ja vastaavat tekniikat, mitkä noudattavat kontekstuaalista tietoa, laatutavoitteita ja budjettirajoituksia.
- Datan kehittämisratkaisujen suunnittelu, missä valitaan tehokkain strategia ja joukko siihen liittyviä tekniikoita ja työkaluja datan laadun parantamiseksi.
- Prosessin uudelleensuunnittelu, missä määritellään prosessin kehittämistoimenpiteet laadun parantamiseksi.
- Parannusten hallinta, missä määritellään uuden organisaatiosäännöt datan laadulle.
- Parannusten seuranta, missä luodaan säännöllisiä seurantatoimia parannusprosessin tuloksien seuraamiseksi. (Batini et al. 2009)

Datan laadun menetelmät omaksuvat kehittämistoimenpiteissä kaksi yleistä strategiatyyppiä, data- ja prosessipohjaisen strategian. Datapohjaiset strategiat parantavat datan laatua suoraan muuttamalla datan arvoja. Esimerkiksi tietokannan vanhentuneita datan arvoja päivitetään ajantasaisemman tietokannan datalla. Prosessipohjaiset strategiat parantavat datan laatua uudelleensuunnittelemalla prosesseja, jotka luovat tai muokkaavat dataa. Esimerkiksi prosessi voidaan suunnitella uudelleen sisällyttämällä siihen toiminto, joka varmistaa datan formaatin ennen sen varastoimista. Datapohjaisissa strategioissa voidaan hyödyntää useita erilaisia tekniikoita. Niitä ovat:

- Uuden datan hankkiminen, mikä parantaa datan laatua korvaamalla huonolaatuista dataa.
- Standardointi (tai normalisointi), joka korvaa tai täydentää epästandardeja arvoja standardin mukaisilla arvoilla. Esimerkiksi lempinimet korvataan oikeilla nimillä tai lyhenteet korvataan koko nimillä.
- Tietueiden linkitys, minkä avulla tunnistetaan dataa, jonka esitykset kahdessa tai useammassa taulussa voivat viitata samaan reaalimaailman kohteeseen.
- Data- ja skeemaintegrointi, mikä määrittelee yhtenäisen näkymän heterogeenisten lähteiden datasta. Integroinnin päätavoitteena on, että käyttäjä voi käyttää heterogeenisten lähteiden dataa yhtenäisen näkymän avulla.
- Lähteiden luotettavuus, missä valitaan datalähteet niiden datan laadun perusteella.
- Virheiden lokalisointi ja korjaus, missä tunnistetaan ja poistetaan datan laatuvirheitä havaitsemalla ne tietueet, jotka eivät täytä laatusääntöjä. Virheiden paikallistamisessa ja korjaamisessa voidaan hyödyntää tilastotiedettä.
- Kustannusten optimointi, mikä määrittelee laadun kehittämistoimenpiteitä minimoimalla kustannuksia. (Batini et al. 2009)

Kaksi prosessipohjaisen strategian päätekniikkaa ovat prosessinhallinta ja prosessien uudelleensuunnittelu. Prosessinhallinta asettaa tarkistuksia ja valvontamenettelyitä, kun uutta dataa luodaan, dataa päivitetään tai prosessin avulla saadaan pääsy uusiin datajoukkoihin. Tämän myötä sovelletaan reaktiivista strategiaa datan muokkaustapahtumiin,

minkä avulla vältetään datan huonontuminen ja virheiden eteneminen. Prosessien uudelleensuunnittelu puolestaan uudistaa prosessit poistaakseen syyt huonolaatuisen datan taustalta ja tarjoaa uusia toimintoja korkealaatuisen datan tuottamiseksi. Jos prosessien uudelleensuunnittelu on radikaalia, tätä tekniikkaa kutsutaan silloin liiketoimintaprosessien uudelleensuunnitteluksi. (Batini et al. 2009)

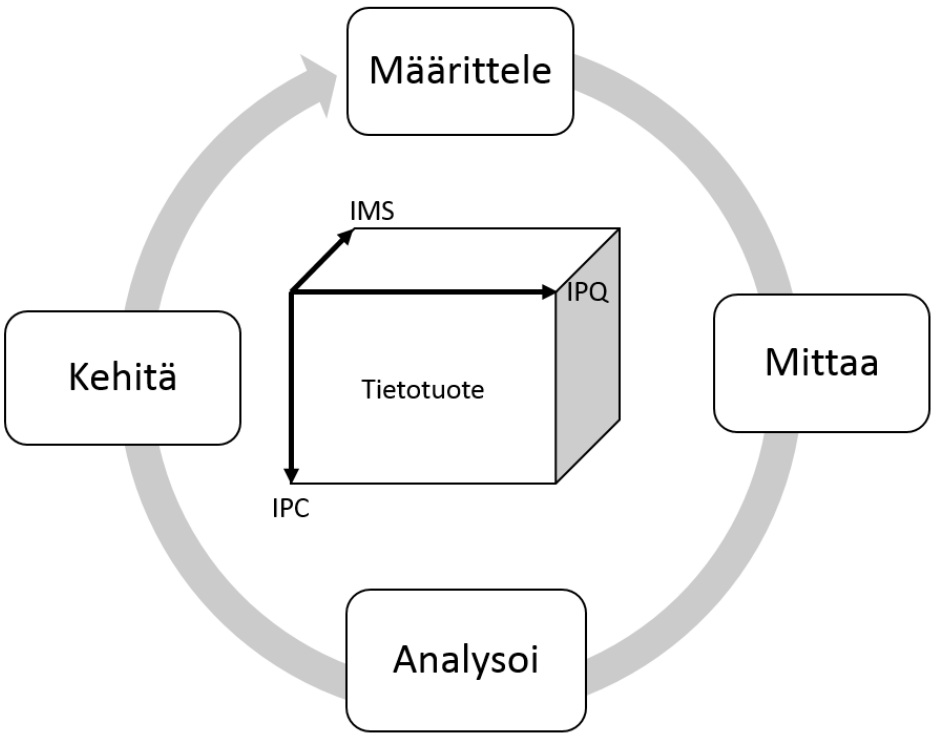
Yleisesti pitkällä aikavälillä prosessipohjaisten tekniikoiden todetaan olevan parempia, koska ne poistavat laatuongelmien juurisyyt. Lyhyen aikavälin näkökulmasta prosessien uudelleensuunnittelu voi olla kuitenkin erittäin kallista. Datapohjaiset strategiat ovat puolestaan kustannustehokkaita lyhyellä aikavälillä, mutta kalliita pitkällä aikavälillä. Ne soveltuvat kertaluontoisiin sovelluskohteisiin, joten niitä suositellaan staattiselle datalle. (Batini et al. 2009)

Kehittyneiden datan laatutekijöiden määrittäminen ei kuitenkaan riitä merkityksellisen datan laadun arvioinnin saavuttamiseksi, koska laadun arviointi on hyvin riippuvainen asiayhteydestä ja sovelluskohteesta. On vaikea saada aikaan yleinen ratkaisu, joka toimii kaikissa tilanteissa. (Andreescu et al. 2014)

4.2 TDQM-menetelmä

Kokonaisvaltainen datan laadunhallinta (*engl. Total Data Quality Management, TDQM*) oli ensimmäinen yleinen menetelmä, joka julkaistiin datan laatukirjallisuudessa. Sitä on käytetty laajasti ohjeena organisatorisen datan uudelleensuunnitteluprojekteissa. Sen perimmäinen tavoite on laajentaa kokonaisvaltaisen laadunhallinnan (TQM) periaatteita datan laatuun. (Batini et al. 2009) Menetelmässä tiedon tuotantojärjestelmällä tarkoitetaan järjestelmää, joka tuottaa tietotuotteita. Tietotuote (*engl. Information product, IP*) korostaa sitä tosiasiaa, että sillä on kuluttajalle siirrettävää arvoa. (Wang 1998)

TDQM:n tavoitteena on tukea datan laadun parantamisprosessia vaatimusanalyysistä toteutukseen. TDQM koostuu neljästä vaiheesta, jotka toteuttavat jatkuvan laadun parantamisprosessin. Vaiheet ovat määritelmä, mittaus, analyysi ja kehittäminen. (Batini et al. 2009) TDQM-syklin määrittelyvaiheessa tunnistetaan tärkeät laadun ulottuvuudet ja vastaavat laadun vaatimukset. Mittausvaiheessa puolestaan tuotetaan laadun mittarit ja analysointivaiheessa tunnistetaan juurisyyt laatuongelmille sekä lasketaan huonon laadun seurauksia. Viimeisessä kehitysvaiheessa esitetään tekniikoita laadun parantamiseksi, mitä sovelletaan laadun ulottuvuuksien ja käyttäjien vaatimusten mukaisesti. (Wang 1998) Kuvassa 7 on esitetty TDQM-syklin vaiheet.



Kuva 7. *TDQM-prosessi (mukaillen Wang 1998)*

Kuvassa esitetyistä lyhenteistä IPC (*engl. Information Product Characteristics*) tarkoittaa tietotuotteen ominaisuuksia, IPQ (*engl. Information Product Quality*) tietotuotteen laatua ja IMS (*engl. Information Manufacturing System*) tiedon tuotantojärjestelmää (Wang 1998). Kokonaisvaltaisessa datan laadunhallinnassa määritellään myös roolit eri vaiheille:

- Tiedon toimittajat, jotka luovat tai keräävät dataa tietotuotteelle.
- Tiedon valmistajat, jotka suunnittelevat, kehittävät tai ylläpitävät dataa ja järjestelmäinfrastruktuuria tietotuotteelle.
- Tiedon kuluttajat, jotka käyttävät tietotuotetta heidän työssään.
- Tietoprosessien johtajat, jotka ovat vastuussa tiedon tuotantoprosessin hallinnasta koko elinkaaren ajan. (Wang 1998; Batini et al. 2009)

Tietotuotteen ominaisuuksia voidaan määritellä kahdella eri tasolla. Korkeammalla tasolla tietotuote käsitteellistetään sen toiminnallisuuksien kannalta tiedon kuluttajia varten. Esimerkiksi asiakastietokannan osalta tiedon kuluttajat tarvitsevat tiettyjä asiakastietoja tehtävien suorittamiseen, kuten asiakasnumeron ja varastotapahtumat. Alemmalla tasolla voidaan tunnistaa tietotuotteen perusyksiköt ja -komponentit sekä niiden suhteet. Perusyksikön määrittäminen on kriittistä, koska se ohjaa tietotuotteen tuottamisen, hyödyntämisen ja hallinnan tavan. Esimerkiksi asiakastietokannassa perusyksikkö olisi ryhmittelemätön asiakastili. (Wang 1998)

Tietotuotteen ominaisuuksien jälkeen tunnistetaan laadun vaatimukset tiedon toimittajien, valmistajien, kuluttajien ja johtajien näkökulmista. Tässä voi hyödyntää esimerkiksi laadun arviointityökalua, jolla kerätään käyttäjiltä tietoa eri ulottuvuuksien laadun tasosta. Tuloksien visualisoinnin myötä on mahdollista tunnistaa tärkeimmät ulottuvuudet ja saada selville eroavaisuuksia käyttäjien tuloksissa. Yhtä tärkeä tehtävä kuin laadun ulottuvuuksien tunnistaminen on myös tiedon tuotantojärjestelmän tunnistaminen. Siinä kuvataan, miten tietotuotteita tuotetaan sekä vuorovaikutukset tiedon toimittajien, valmistajien, kuluttajien ja IP-johtajien kanssa. (Wang 1998)

Mittaamisvaiheen ydin on laadun mittareiden kehittäminen. Asiakastietokannan tapauksessa mittareita voidaan suunnitella seuraamaan esimerkiksi asiakkaiden virheellisten postinumeroiden prosenttiosuuksia, asiakastietojen päivitystiheyttä sekä niiden asiakastilien määrää, joita ei ole olemassa. (Wang 1998)

Monimutkaisemmalla tasolla on otettava huomioon myös yrityksen liiketoimintasäännöt. Mittareita voidaan kehittää myös tiedon valmistuksen näkökulmasta. IP-tiimi voi haluta seurata esimerkiksi:

- Mikä osasto teki suurimman osan päivityksistä järjestelmään viime viikolla?
- Kuinka monta luvatonta sisäänkirjautumista on ollut?
- Kuka keräsi raakadatan asiakastilille? (Wang 1998)

Mittaustulosten perusteella IP-tiimi tutkii laatuongelmien juurisyyt. Viimeisessä eli kehittämisvaiheessa tunnistetaan kehityksen avainalueet, kuten esimerkiksi tieto- ja työvirtojen linjaaminen vastaavan tiedon valmistusjärjestelmän kanssa tai tietotuotteen keskeisten ominaispiirteiden uudelleensuuntaaminen yrityksen tarpeiden mukaisesti. (Wang 1998)

TDQM-menetelmässä annetaan myös ohjeita sen hyödyntämiseksi. Organisaation täytyy:

(1) Ilmaista selkeästi tietotuotteet liiketoiminnan termein
(2) Perustaa IP-tiimi, joka koostuu ylemmästä TDQM:n johtajasta, menetelmään perehtyneestä IP-insinööristä sekä jäsenistä, jotka ovat tiedon toimittajia, valmistajia, kuluttajia ja IP-johtajia
(3) Opettaa laadun arviointia ja hallintaa kaikille IP-osallisille
(4) Vakiinnuttaa jatkuva IP-kehitys. (Wang 1998; Batini et al. 2009)

4.3 AIMQ-menetelmä

AIMQ-menetelmä (*engl. A Methodology for Information Quality Assessment, AIMQ*) on vertailuanalyysiin keskittyvä datan laadun menetelmä (Batini et al. 2009). Menetelmän ensimmäinen komponentti on 2x2-malli tai -kehys, jota kutsutaan myös PSP/IQ-malliksi. Sen avulla selvitetään laadun merkityksiä kuluttajille ja johtajille. Mallissa on neljä

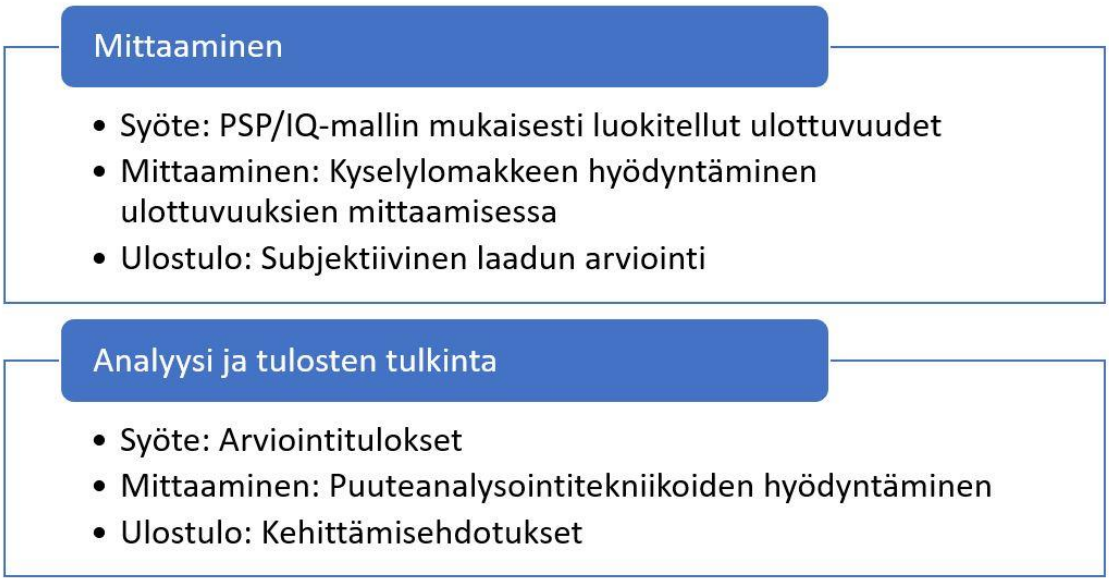
kvadranttia riippuen siitä, pidetäänkö dataa tuotteena vai palveluna ja arvioidaanko kehitystoimenpiteitä virallisen määrittelyn vai asiakkaiden odotusten perusteella. (Lee et al. 2002) Taulukossa 7 on esitetty PSP/IQ-mallin sisältö.

Taulukko 7. *PSP/IQ-malli (mukaillen Lee et al. 2002)*

	Noudattaa määrittelyjä	**Vastaa kuluttajien odotuksia**
Tuotteen laatu	Laadukas data Oikeellisuus Täydellisyys Esityksen ytimekkyys Esityksen johdonmukaisuus	Hyödyllinen data Sopiva määrä Asianmukaisuus Ymmärrettävyys Tulkittavuus Objektiivisuus
Palvelun laatu	Luotettava data Oikea-aikaisuus Turvallisuus	Käytettävä data Uskottavuus Saatavuus Helppokäyttöisyys Maine

Toinen komponentti on kyselylomake, jolla mitataan datan laatua käyttäjille ja johtajille tärkeiden ulottuvuuksien mukaan (Lee et al. 2002). Ensimmäistä pilottikyselyä käytetään tunnistamaan asiaankuuluvat laadun ulottuvuudet ja attribuutit vertailuanalyysiä varten. Toisessa kyselyssä käsitellään aiemmin tunnistettuja ulottuvuuksia ja attribuutteja laadun mittauksien saamiseksi. Lopuksi näitä mittauksia verrataan vertailuarvoihin. (Batini et al. 2009) Liitteessä B on esitetty AIMQ-kyselylomake.

Menetelmän kolmas osa koostuu kahdesta puuteanalyysiin (*engl. Gap Analysis*) liittyvästä tekniikasta kyselylomakkeen tuloksien tulkitsemiseksi. Ensimmäinen tekniikka (*engl. Benchmarking Gap Analysis*) vertaa organisaation laatuarvoja toiseen organisaatioon, joka hyödyntää alan parhaita käytäntöjä. Toinen tekniikka (*engl. Role Gap Analysis*) puolestaan mittaa eri sidosryhmien arviointien etäisyyksiä. Sen tarkoituksena on paljastaa poikkeamia eri roolien arviointien välillä mahdollisten laatuongelmien havaitsemiseksi. (Lee et al. 2002) Kuvassa 8 on esitetty yhteenveto AIMQ-menetelmän vaiheista.



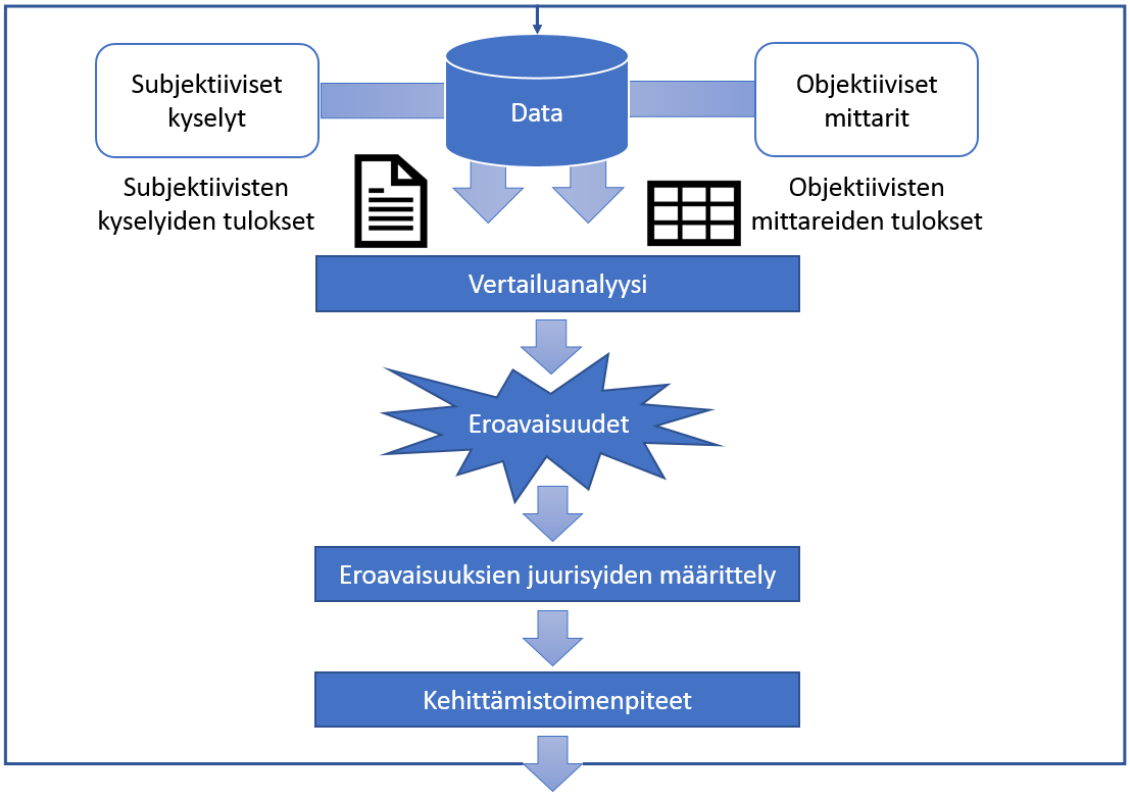
Kuva 8. *AIMQ-menetelmän vaiheet (mukaillen Batini 2009)*

AIMQ-menetelmän jokaisella komponentilla on arvoa. Menetelmän keskeinen myötävaikutus ilmentyy kuitenkin näiden komponenttien integroinnista ja synteesistä. Oikein sovellettuna ne muodostavat yhdessä tehokkaan menetelmän laadun arvioimiseksi erilaisissa organisaatioissa, joissa on tehtävä päätöksiä tehtävien priorisoimisesta ja resurssien allokoimisesta laadun kehittämiseksi. (Lee et al. 2002)

4.4 DQA-menetelmä

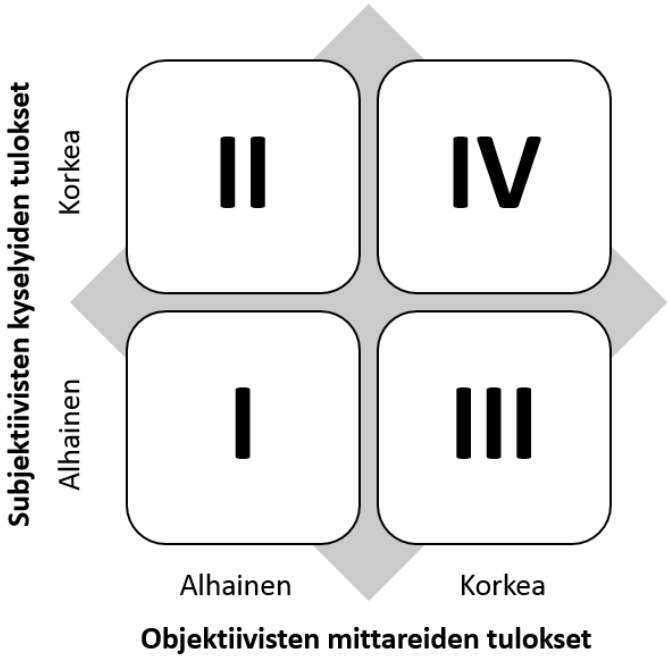
DQA-arviointimenetelmä (*engl. Data Quality Assessment, DQA*) on subjektiiviset ja objektiiviset mittaukset yhdistävä lähestymistapa. Menetelmän käyttäminen organisaation datan laadun kehittämiseksi vaatii kolme vaihetta, joita on havainnollistettu kuvassa 9:

- Subjektiivisten ja objektiivisten datan laadun mittauksien toteuttaminen.
- Tuloksien vertaileminen, eroavuuksien tunnistaminen ja juurisyiden määrittäminen.
- Tarvittavien kehitystoimenpiteiden määrittäminen ja toteuttaminen. (Pipino et al. 2002)



Kuva 9. *DQA-menetelmä (mukaillen Pipino et al. 2002)*

Analyysin aluksi verrataan tietyn ulottuvuuden subjektiivisia ja objektiivisia mittauksia. Analyysin tulokset sijoitetaan yhteen neljästä kvadranteista, mitkä ovat esitetty kuvassa 10. Tämän lähestymistavan tavoitteena on saavuttaa kvadrantin IV laadun tila, jolloin subjektiivisten ja objektiivisten mittauksien tulokset viittaavat korkeaan datan laatuun. Jos taas analyysi osoittaa kvadrantteja I, II tai III, niin yrityksen on tutkittava juurisyyt ja ryhdyttävä korjaaviin toimenpiteisiin. Korjaustoimenpiteet ovat kuitenkin tapauskohtaisia. (Pipino et al. 2002)



Kuva 10. *Subjektiivisten ja objektiivisten mittauksien tuloskvadrantti (mukaillen Pipino et al. 2002)*

Esimerkkiyrityksen (GCG) tapauksessa subjektiivisten kyselyiden tulokset ryhmien välillä osoittivat, että johdonmukaisuus ja täydellisyys olivat kaksi suurinta huolenaihetta. Kun näitä arvioita verrattiin objektiivisiin mittauksiin, niiden huomattiin tukevan subjektiivisia kyselyitä. Tällöin tulokset viittasivat kvadrantin I laatuun. Tämä yhteisymmärrys johti siihen, että kyseisessä yrityksessä tehtiin laaja aloite datan johdonmukaisuuden ja täydellisyyden parantamiseksi. (Pipino et al. 2002)

Datan laadun parantamisen tavoittelussa ei voida hyödyntää joukkoa mittareita, jotka sopivat kaikkiin tilanteisiin. Datan laadun arvioiminen on jatkuvaa työtä, joka edellyttää tietoisuutta subjektiivisten ja objektiivisten datan laatumittareiden kehittämisen perusperiaatteista. (Pipino et al. 2002)

4.5 Hybridi-menetelmä

Monia erilaisia arviointitekniikoita on kehitetty, mutta ei ole kuitenkaan selvää, onko niitä tarpeen kehittää vielä enemmän eri asiayhteyksiin. Teoreettisesti yleinen arviointimenetelmä saattaa olla olemassa, ja olemassa olevia arviointitekniikoita voidaan pitää jossain määrin sen ilmentyminä. (Woodall et al. 2013)

Hybridi-menetelmän tarkoituksena on osoittaa, miten uusia arviointitekniikoita voidaan kehittää yhdistelemällä olemassa olevien arviointitekniikoiden toimintoja. Tämä lähestymistapa pyrkii siten myös välttämään tarpeettomien toimintojen suorittamista. Keskeistä menetelmässä on, että voidaan ottaa toiminoja yhdestä arviointitekniikasta ja integroida

ne toisten arviointitekniikoiden kanssa täysin kustomoidun arviointitekniikan muodostamiseksi. (Woodall et al. 2013)

Hybridi-menetelmä koostuu neljästä vaiheesta, joiden avulla kehitetään uusia arviointitekniikoita tiettyihin organisatorisiin vaatimuksiin. Ensimmäinen vaihe on arvioinnin tavoitteen määrittely. Tavoitteena voi olla esimerkiksi:

- Mitata tiettyä tunnistettua datan laatuongelmaa
- Määritellä ja priorisoida organisaation datan laatuongelmia ja saada niille mittarit. (Woodall et al. 2013)

Toinen vaihe on yrityksen datan laadun arviointiin liittyvien vaatimusten tunnistaminen. Eri yrityksellä on erilaisia vaatimuksia datan laadun arviointiin. Tämä vaihe edellyttää, että organisaatio haluaa arvioida datan laatua ja tunnistaa siihen liittyvät vaatimukset. Vaatimusten asianmukaisuuden vuoksi on hyvä tarkistaa, että jokainen vaatimus noudattaa ensimmäisessä vaiheessa määriteltyä tavoitetta ja edistää sen saavuttamista. Vaatimuksia voivat olla esimerkiksi:

- Huonon datan aiheuttamien kustannusten määrittäminen
- Alustavan kustannusarvion hankkiminen arvioinnin resurssien perustelemiseksi
- Datan luomisen ja virtaamisen mallintaminen
- Olemassa olevien datamallien kerääminen. (Woodall et al. 2013)

Kolmas vaihe on niiden arviointitoimintojen valitseminen, jotka kohtaavat organisatoristen tavoitteiden kanssa. Tämän vaiheen tarkoituksena on valita taulukon 8 luettelosta tavoitteita vastaavat toiminnot. Käytännössä toisen vaiheen vaatimusten tunnistaminen voidaan tehdä näistä erilaisista toiminnoista saatavan ymmärryksen avulla. Tämän vuoksi iteratiivinen prosessi on todennäköinen, jossa toista ja kolmatta vaihetta käytetään täydentämään toisiaan ja mahdollistamaan tarvittavien vaatimusten muodostaminen. (Woodall et al. 2013)

Taulukko 8. *Arviointitekniikoihin liittyviä toimintoja (mukaillen Woodall et al. 2013)*

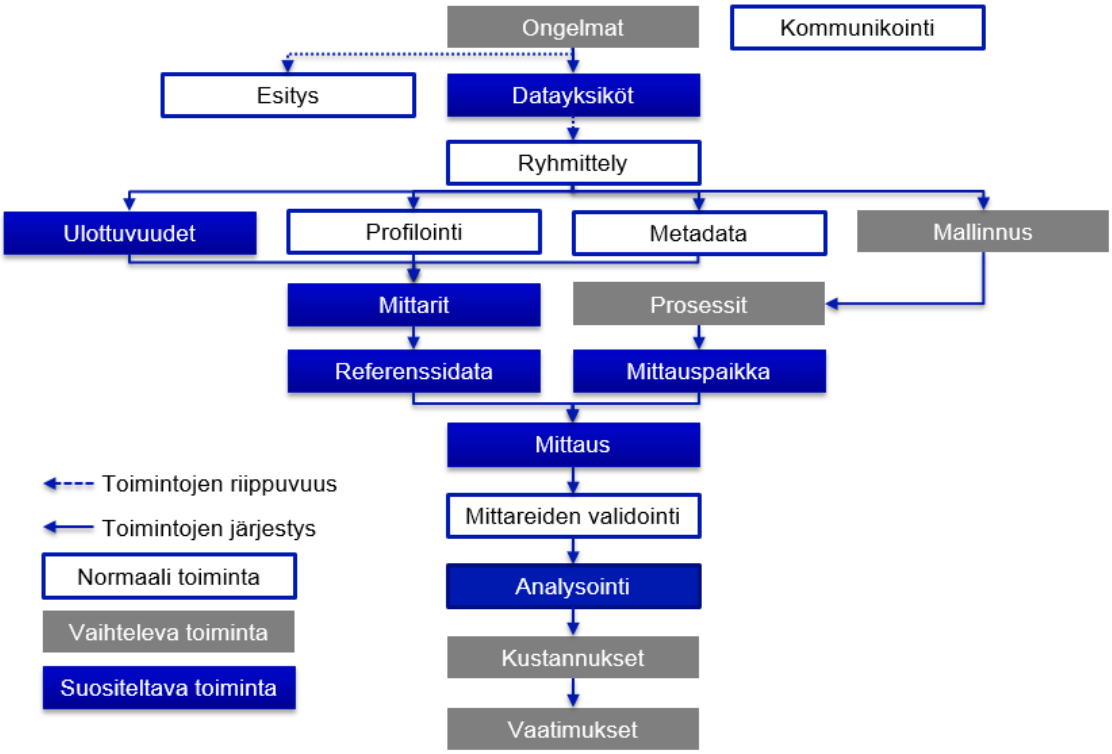
Toiminto ja lyhenne	Määritelmä
Arviointiprojektin esittäminen ylimmälle johdolle (esitys)	Tarkoituksena on saada ylimmän johdon tuki datan laadun arviointiprojektille.
Datan luomisen ja virtaamisen mallintaminen (mallinnus)	Luodaan malli, miten dataa luodaan, päivitetään, poistetaan ja siirretään yhdestä lähteestä toiseen.
Datan profiloinnin suorittaminen (profilointi)	Prosessi, jossa tutkitaan dataa ja kerätään siitä tilastotietoja.

Datayksiköiden ryhmittely (ryhmittely)	Prosessi, jossa ryhmitellään datayksiköitä eri luokkiin. Ryhmittelykriteereitä voivat olla esimerkiksi datan tyyppi ja riskitaso.
Datayksiköiden valitseminen (datayksiköt)	Prosessi, jossa valitaan arvioinnin kohteeksi datan arvot, attribuutit, taulukot, tietojärjestelmät jne. Tähän voi sisältyä myös datan keräysprosessi tarvittavien datan arvojen saamiseksi.
Kustannusten tunnistaminen (kustannukset)	Prosessi, jossa määritetään alhaisen datan laadun aiheuttamat liiketoiminnalliset vaikutukset ja/tai taloudelliset tappiot.
Metadatan kerääminen (metadata)	Prosessi, jossa kerätään asiaankuuluvat metadatat, kuten datamallit.
Mittareiden tunnistaminen (mittarit)	Prosessi, jossa tunnistetaan, kehitetään tai käytetään olemassa olevia datan laadun mittareita.
Mittareiden validointi	Prosessi, jossa tarkistetaan datan laadun mittareiden ja niiden toteutuksen korrektius.
Mittauspaikan valitseminen (mittauspaikka)	Määritetään missä ja milloin dataa mitataan. Voidaan täsmentää myös subjektiivisten mielipiteiden antajat.
Objektiivisten/subjektiivisten mittauksien suorittaminen (mittaus)	Menetelmä objektiivisten mittauksien ja/tai subjektiivisten mielipiteiden saamiseksi.
Organisaation ongelmien tunnistaminen ja priorisointi (ongelmat)	Luettelo datan laatuun liittyvistä ongelmista.
Prosessien valitseminen (prosessit)	Valitaan ne liiketoimintaprosessit, joihin keskitytään arvioinnissa.
Referenssidatan tunnistaminen (referenssidata)	Määritetään vertailudata, jota voidaan käyttää syötteenä valittuihin mittareihin.
Tulosten analysointi (analysointi)	Datan laadun mittauksien arvojen analysointi.
Tulosten kommunikointi ja jakaminen (kommunikointi)	Datan laadun arvioinnin tulosten tai nykytilanteen kommunikointi ja jakaminen asiaankuuluvien ihmisten kanssa.
Ulottuvuuksien tunnistaminen (ulottuvuudet)	Prosessi, jossa tunnistetaan ulottuvuudet tai hyödynnetään olemassa olevaa mallia ulottuvuuksista, kuten PSP/IQ.
Vaatimusten määrittäminen (vaatimukset)	Prosessi, jossa määritetään datan laadun vaatimustaso. Vaatimuksia voidaan verrata mittausarvoihin tarvittavan laadun parannustason määrittämiseksi.

Neljäs eli viimeinen vaihe on arviointitoimintojen konfigurointi. Tämän vaiheen tarkoituksena on järjestää toiminnot järkevään järjestykseen ja sisällyttää toimintojen riippuvuussuhteet. Suositellut toiminnot ovat niitä, joita suositellaan ottamaan mukaan kaikkiin uusiin arviointitekniikkoihin. Niitä tulisi poistaa vain, jos on perusteltu syy olla suorittamatta niitä. Suositeltuja toimintoja ovat:

(1) Datayksiköiden valitseminen
(2) Mittauspaikan valitseminen
(3) Referenssidatan tunnistaminen
(4) Datan laadun ulottuvuuksien tunnistaminen
(5) Datan laadun mittareiden tunnistaminen
(6) Mittauksen suorittaminen
(7) Tulosten analysointi

Mittausprosessista (6) saadaan arvot ulottuvuuksille (4) ja mittareille (5) tiettyjen datayksiköiden (1) avulla. Mittauksista saatavat arvot ovat hyödyttömiä, kunnes tulosten analysointia (7) sovelletaan. Datan mittauspaikka on tunnistettava (2), jotta tiedetään, missä mittausprosessin kohdassa mittareita hyödynnetään. Referenssidatan tunnistaminen (3) on ehdollinen riippuen siitä, mitä mitataan. Esimerkiksi referenssidataa voidaan tarvita syötteenä mittarille, joka mittaa tarkkuutta. (Woodall et al. 2013) Kuvassa 11 on esitetty yleiseen arviointitekniikkaan kuuluvat toiminnat. Kyseisessä kuvassa siniset laatikot osoittavat suositeltavia toimintoja, harmaat laatikot vaihtelevia toimintoja, joita voidaan sijoittaa eri kohtiin, sinireunaiset laatikot normaaleja toimintoja, nuolet ilmaisevat toimintojen järjestystä ja katkoviivat viittaavat toimintojen välisiin riippuvuuksiin.



Kuva 11. *Yleisen arviointitekniikan toimintoja lyhenteiden avulla esitettynä (mukaillen Woodall et al. 2013)*

Kommunikointi-toiminnolla ei ole linkkejä muihin toimintoihin ja sitä voidaan tehdä missä tahansa arvioinnin kohdassa. Sitä suositellaan tehtävän useissa kohdissa korkeatasoisen viestinnän ylläpitämiseksi ulkopuolisille sidosryhmille. Datan laadun profilointia voidaan suorittaa mittareiden määrittämisen jälkeen, jolloin profilointiohjelmisto laskee automaattisesti arvot mittareille. Profilointia voidaan hyödyntää myös sopivien mittareiden kehittämisessä, eli ennen mittareiden määrittämistä. (Woodall et al. 2013)

Kuvan 11 yleinen arviointitekniikka ei kuitenkaan tarkoita ”vesiputous”-tyylisen prosessin suorittamista, sillä suositeltavampaa on toteuttaa iteraatioita toimintojen välillä. Esimerkiksi ulottuvuuksien tunnistamisessa saattaa korostua tarve valittujen datayksiköiden uudelleentarkistukseen, minkä myötä iteraatioita voidaan tehdä useitakin kertoja. (Woodall et al. 2013)

Hybridi-menetelmää voidaan pitää yhtenä askeleena kohti datan laadun arvioinnin edellyttämien perustoimien tunnistamista. Tätä voidaan pitää tärkeänä, koska tiettyyn sovelluskohteeseen sopivan arviointitekniikan kehittäminen on helppoa, mutta ei ole kuitenkaan helppoa tunnistaa, tarvitaanko uusi arviointitekniikka uudelle sovelluskohteelle ja mitä komponentteja siihen tarvitaan. Useita arviointitekniikoita on kehitetty, mitkä antavat oman näkökulmansa datan laadun arvioimiseen. Monien arviointitekniikoiden tuomat vaihtoehdot voivat olla hyödyllisiä, mutta useimmiten ne eivät eroa toisistaan merkittävästi. Hybridi-menetelmän tarjoamaa mallia arviointitekniikoiden yleisistä toiminnoista

voidaan hyödyntää esimerkiksi tunnistamaan eri arviointitekniikoiden eroavaisuuksia. (Woodall et al. 2013)

4.6 Datan laadun mittaamisen ja arvioinnin viitekehys

Tässä alaluvussa esitetään teorian yhteenveto, jota voidaan pitää myös aiheen viitekehyksenä. Yhteenveto on jaoteltu teorialukujen mukaisesti datan laatuun, datan laadun mittaamiseen ja datan laadun arviointiin.

DATAN LAATU

Datan tyypit voidaan jaotella rakenteelliseen, rakenteettomaan ja puolirakenteelliseen dataan (Batini et al. 2009; Aljumaili et al. 2016). Datan laadun kirjallisuudessa keskitytään pääosin rakenteelliseen dataan (Batini et al. 2009). Dataa voidaan myös laajemmin luokitella Master dataan, transaktiodataan, referenssidataan, metadataan, historiadataan ja väliaikaiseen dataan. Referenssidataa tarvitaan Master datan luomiseksi, Master dataa tarvitaan transaktiodatan luomiseksi ja metadataa tarvitaan muiden datan kategorioiden ymmärtämiseen. (McGilvray 2008 ss. 42-44)

Datan laadulla tarkoitetaan sitä, miten hyvin se sopii datan kuluttajien käyttötarpeisiin (Wang & Strong 1996; Watts et al. 2009). Keskeistä on myös huomioida, miten hyvin data esittää kuluttajien mielestä sitä, mitä se on tarkoituskin esittää (Sebastian-Coleman 2013 s. 40). Datan laatua voidaan tarkastella esimerkiksi suunnittelun laadun ja vaatimustenmukaisuuden laadun näkökulmista (Heinrich et al. 2009). Datan laatuongelmat voivat johtua ihmisten, prosessien tai järjestelmien ongelmista (McGilvray 2008 s. 5).

Datan laadun ulottuvuuksilla tarkoitetaan laatuattribuutteja, jotka esittävät yhtä datan laadun näkökulmaa (Wang & Strong 1996). Ne ovat yleisesti mitattavia kategorioita tiettyjen datan ominaisuuksien mukaan (Sebastian-Coleman 2013 s. 40). Laadun ulottuvuuksia voidaan luokitella esimerkiksi luontaiseen laatuun, kontekstuaaliseen laatuun, esitystavan laatuun ja saavutettavuuden laatuun (Wang & Strong 1996; Lee et al. 2002). Jokainen datan laadun ulottuvuus vaatii erilaisia työkaluja, tekniikoita ja prosesseja sen mittaamiseksi (McGilvray 2008 ss. 30-31).

Datan laadun kustannukset ovat laadun arvioinnin ja parannustoimien kustannusten summa, mitä kutsutaan myös datan laatuohjelman kustannuksiksi (Batini et al. 2009). Datan laadun parantamisen kustannuksiin kuuluvat esimerkiksi ennaltaehkäisy-, selvitys- ja korjauskustannukset (Eppler & Helfert 2004). Huonolaatuisen datan kustannuksia voidaan luokitella prosessikustannuksiin ja vaihtoehtokustannuksiin tai suoriin ja epäsuoriin kustannuksiin (Eppler & Helfert 2004; Batini et al. 2009).

DATAN LAADUN MITTAAMINEN

Datan laadun mittaamisessa on otettava huomioon erilaiset ulottuvuudet ja kehitettävä mittausmenetelmät valituille ulottuvuuksille (Wang et al. 1995; Wang & Strong 1996; Bronselaer et al. 2018b). Data on aineetonta, mutta sitä luodaan ja tallennetaan asiayhteydessä, mikä mahdollistaa sen mittaamisen (Sebastian-Coleman 2013 ss. 42-53). Datan laadun mittaamisella tarkoitetaan toimintoa, jossa määritetään numeroarvo tarkastelun kohteena olevalle attribuutille (Caballero et al. 2007). Mittauksista saadaan arvoja, joita arvioinnissa tarkastellaan tarvittavien datan laadun kehittämistoimenpiteiden määrittämiseksi (Woodall et al. 2013).

Datan laadun mittaaminen voi olla objektiivista tai subjektiivista (Pipino et al. 2002; Batini et al. 2009; Sebastian-Coleman 2013 s. 60; Bronselaer et al. 2018a). Mittauksia voidaan jakaa samojen periaatteiden mukaisesti myös rakenne- ja sisältöpohjaisiin mittauksiin. Objektiivinen (rakennepohjainen) mittaaminen perustuu datan fyysisiin ominaisuuksiin, kuten lukumäärien suhteisiin tai aikamittauksiin. (Ballou & Pazer 2003; Even & Shankaranarayanan 2005; Even & Shankaranarayanan 2009; Watts et al. 2009; Aljumaili et al. 2016; Bronselaer et al. 2018a) Objektiiviset mittarit voivat olla tehtävästä riippumattomia (kuvaavat datan tilaa ilman asiayhteystietoa) tai tehtävästä riippuvaisia (sisältävät esim. yrityksen liiketoimintasäännöt) (Pipino et al. 2002). Subjektiivinen (sisältöpohjainen) mittaaminen puolestaan perustuu datan käyttäjien mielipiteisiin ja se heijastaa käyttäjien tarpeita ja odotuksia (Batini et al. 2009; Sebastian-Coleman 2013 s. 60). Taloudellisesta näkökulmasta objektiivinen mittaaminen voidaan liittää kustannuksiin ja subjektiivinen mittaaminen datan laadun parantamisesta saataviin hyötyihin (Even & Shankaranarayanan 2009).

Mittauksien on oltava ymmärrettäviä, minkä vuoksi ne vaativat metadataa. Muita vaatimuksia mittaamiselle ovat ymmärrys datan liiketoimintakäsitteistä, liiketoiminnallisista ja teknisistä dataprosesseista sekä datamalleista. (Sebastian-Coleman 2013 ss. 44-65) Tarkempia vaatimuksia itse mittareille ovat minimi- ja maksimiarvojen olemassaolo, välimatka-asteikolliset arvot, konfigurointiparametrien ja mittarien määrittämisen laatu, mittarien arvojen yhdistäminen sekä mittarin taloudellinen tehokkuus (Heinrich et al. 2018a).

Datan profilointi on prosessi datajoukkojen ominaisuuksien analysoimiseen, mistä saadaan tulokseksi tilastollisia indikaattoreita ja metadataa (Andreescu et al. 2014). Profiloinnin keskeisimmät menetelmät voidaan jakaa rakenteen, sisällön ja suhteiden analysointiin (Dorr & Murnane 2011; Andreescu et al. 2014; Mahanti 2014; Azeroua et al. 2018). Profiloinnin tehtäviä voidaan jaotella tarkemmin myös metadatan, esityksen, sisällön, dataryhmien ja loogisten sääntöjen profilointiin (Dai et al. 2016).

Datan laadun mittaamisen menetelmissä on joustavuutta, sillä jokaista ulottuvuutta voidaan mitata useilla eri tavoilla (Aljumaili et al. 2016). Mittareiden valinnassa voidaan huomioida useita eri tekijöitä, kuten mittarin prioriteetti, mittausmenetelmä, mittaustiheys, kustannusten ja hyötyjen suhde sekä huomiotta jättämisen riski (Umar et al. 1999).

Yleensä datan laadun mittareiden arvot vaihtelevat välillä 0-1, jossa 1 tarkoittaa toivottua datan laatua ja 0 viittaa vähiten toivottuun laatuun (Pipino et al. 2002; Cappiello et al. 2004; Even & Shankaranarayanan 2009; Blake & Mangiameli 2011; Aljumaili et al. 2016). Tyypillisissä mittareissa luvusta yksi vähennetään epämieluisten tulosten lukumäärä jaettuna kokonaislukumäärällä (Pipino et al. 2002; Caballero et al. 2007; Aljumaili et al. 2016). Kaikille datan laadun ulottuvuuksille (esim. asiaankuuluvuus ja tulkittavuus) ei voi suoraan määrittää määrällistä mittaria, vaan ne vaativat käyttäjäkyselyiden toteuttamista (Cappiello et al. 2004; Watts et al. 2009).

Datan laadun mittaamisen työkalulle tai ympäristölle voidaan määrittää viisi vaatimusta, jotka ovat mukautettavat laatumittarit, datan laadun yleiskuvaus, virheilmoitukset, virhejakauma ja datan tutkiminen (Bors et al. 2018). Mittareiden esittämisessä voidaan huomioida myös eri yleisöjen tarpeet, minkä myötä mittareita voidaan esittää yhteenveto- ja porautumistasossa sekä yksityiskohtaisissa raporteissa (McGilvray 2008 s. 270).

DATAN LAADUN ARVIOINTI

Datan laadun arvioinnilla tarkoitetaan prosessia, jossa hyödynnetään datan laadun mittauksia laadun diagnosoimiseksi ja tarvittavien datan laadun kehittämistoimenpiteiden määrittämiseksi (Woodall et al. 2013). Mittauksen keskeisimpänä tarkoituksena on määrällisen merkityksen tarjoaminen siitä, kuinka monessa laadun ulottuvuudessa päästään tavoitteeseen (Caballero et al. 2007). Datan laadun arviointimenetelmien analysoimiseen ja vertailemiseen on olemassa useita eri näkökulmia. Yleisimmissä tapauksissa arviointimenetelmät koostuvat kolmesta vaiheesta, jotka ovat tilan rekonstruktio, mittaus sekä kehittämistoimenpiteet. (Batini et al. 2009)

TDQM-arviointimenetelmä oli ensimmäinen yleinen menetelmä, joka julkaistiin datan laatukirjallisuudessa. Sen tavoite on laajentaa kokonaisvaltaisen laadunhallinnan (TQM) periaatteita datan laatuun. (Batini et al. 2009) TDQM koostuu neljästä vaiheesta, jotka ovat määritelmä, mittaus, analyysi ja kehittäminen, ja siinä määritetään myös roolit näille vaiheille. TDQM-menetelmässä annetaan myös nelivaiheinen ohjelista sen hyödyntämiseen. (Wang 1998; Batini et al. 2009)

AIMQ-arviointimenetelmä on vertailuanalyysiin keskittyvä menetelmä (Batini et al. 2009). Sen ensimmäinen komponentti on 2x2-malli (PSP/IQ-malli), jonka avulla selvitetään laadun merkityksiä kuluttajille ja johtajille. Menetelmän toinen komponentti on kyselylomake, jolla mitataan datan laatua käyttäjille ja johtajille tärkeiden ulottuvuuksien mukaan. AIMQ-menetelmän kolmas osa koostuu kahdesta puuteanalyysiin liittyvästä tekniikasta kyselylomakkeen tuloksien tulkitsemiseksi. (Lee et al. 2002)

DQA-arviointimenetelmä yhdistää subjektiiviset laadulliset mittaukset ja objektiiviset määrälliset mittaukset. Kyseisten mittauksien lisäksi menetelmään kuuluvat tuloksien

vertaileminen, eroavaisuuksien tunnistaminen ja juurisyiden määrittäminen sekä tarvittavien kehitystoimenpiteiden toteuttaminen. Mittauksien tuloksien analysoimisessa hyödynnetään tuloskvadranttia, jossa tavoitteena on saavuttaa kvadrantin IV laadun tila, jolloin subjektiivisten ja objektiivisten mittauksien tulokset viittaavat korkeaan datan laatuun. (Pipino et al. 2002)

Hybridi-arviointimenetelmän tarkoituksena on osoittaa, miten uusia arviointitekniikoita voidaan kehittää yhdistelemällä olemassa olevien arviointitekniikoiden toimintoja. Menetelmä koostuu neljästä vaiheesta, jotka ovat arvioinnin tavoitteen määrittely, vaatimusten tunnistaminen, arviointitoimintojen valitseminen sekä arviointitoimintojen konfigurointi. Menetelmään liittyy lista arviointitekniikoihin liittyvistä toiminnoista, joista valitaan tavoitteita vastaavat toiminnot. (Woodall et al. 2013)

5. TUTKIMUKSEN TOTEUTTAMINEN

Tutkimuksen empiria toteutetaan Hybridi-arviointimenetelmän mukaisesti. Kyseinen menetelmä valittiin, koska sen avulla voidaan luoda täysin kustomoitu arviointiprojekti kohdeyrityksen vaatimusten ja tavoitteiden mukaisesti. Siihen sisällytetään myös muita arviointi- ja mittausmenetelmiä, minkä myötä se kokoaa työssä käsiteltyjä aiheita yhteen. Mittauksissa huomioitiin sekä objektiiviset määrälliset mittaukset, että subjektiiviset laadulliset mittaukset, minkä vuoksi työssä hyödynnetään määrällistä ja laadullista aineistoa. Tämä luku on jaoteltu Hybridi-arviointimenetelmän vaiheiden mukaisesti arvioinnin tavoitteen määrittämiseen, arvioinnin vaatimusten tunnistamiseen, arviointitoimintojen valitsemiseen ja arviointitoimintojen konfigurointiin. Ennen Hybridi-menetelmän läpikäymistä tuodaan esiin tiivistetysti kohdeyrityksen keskeisiä piirteitä.

5.1 Kohdeyritys

Työn kohdeyrityksenä on suomalainen ICT-alan yritys, jonka henkilöasiakasliiketoiminta on tarkastelun kohteena. Diplomityön toteuttamishetkellä yrityksen Master data oli vielä kehitysvaiheessa, minkä vuoksi sitä ei hyödynnetty koko yrityksen tasoisesti. Master datan hyödyntämisen myötä pystyttiin kuitenkin haastattelemaan sen käyttäjiä ja saamaan mielipiteitä laadun tasosta.

Kohdeyritys on kooltaan suuryritys ja sen suuri data- ja työntekijämäärä sekä liiketoimintojen laajuus tuovat omat haasteensa datan laadun kehittämiseen. Eri liiketoimintayksiköillä ja työntekijöillä saattavat olla omat menetelmät datan laadun tarkasteluun, minkä myötä yhdenmukainen näkymä puuttuu. Lisäksi samasta datasta voi esiintyä useita erilaisia muunnoksia, jotka tuovat haasteita esimerkiksi attribuuttien merkityksien ymmärtämiseen.

Diplomityön tekijä on työskennellyt kohdeyrityksessä ennen diplomityön aloittamista Master datan hallinnan parissa. Työtehtäviin kuului myös datan laadun tarkastelu, mikä pohjautui enimmäkseen yrityksen liiketoimintasääntöjen hyödyntämiseen. Yrityksessä nousi esiin tarve laajemmalle datan laadun ja sen menetelmien tarkastelulle, mikä lopulta päätyi diplomityön aiheeksi.

5.2 Arvioinnin tavoitteen määrittäminen

Arvioinnin tavoitteena on mitata tiettyjä tunnistettuja datan laatuongelmia objektiivisten ja subjektiivisten mittauksien avulla. Objektiiviset mittarit perustuvat määrällisiin matemaattisiin kaavoihin, kun taas subjektiiviset mittaukset ovat datan käyttäjien mielipiteitä laadun ulottuvuuksista.

Toisena keskeisenä tavoitteena on toteuttaa datan laadun arviointia ja mittaamista jatkuvan parantamisen periaatteiden mukaisesti. Tarkoittaen sitä, että mittaamisen ja arvioinnin menetelmiä voidaan hyödyntää eri datoihin tietyin aikavälein, mikä mahdollistaa suuntauksien seuraamisen.

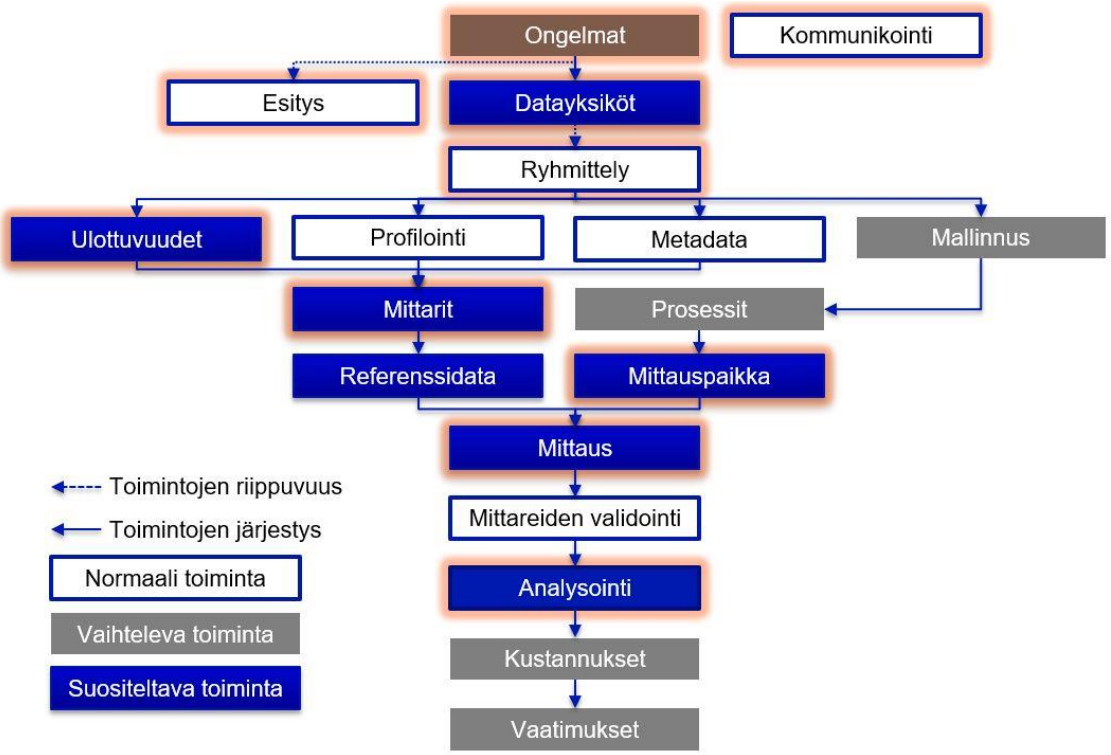
5.3 Arvioinnin vaatimusten tunnistaminen

Datan laadun arvioinnin ensisijaisena vaatimuksena voidaan pitää datan laadun, ulottuvuuksien, mittaamisen sekä arvioinnin ymmärtämistä. Keskeisten asioiden tunnistaminen ja ymmärtäminen mahdollistaa perustavanlaatuisen arviointiprosessin toteuttamisen. Muita kohdeyrityksen kanssa tunnistettuja vaatimuksia olivat:

- Datan laatuongelmien tunnistaminen
- Datan laadun ulottuvuuksien tunnistaminen, priorisointi ja valitseminen
- Objektiivisten ja subjektiivisten mittauksien hyödyntäminen
- Tuloksien analysointi

5.4 Arviointitoimintojen valitseminen

Taulukon 8 luettelosta valittiin kohdeyrityksen tavoitteita ja vaatimuksia vastaavat toiminnot. Kuvassa 12 on esitetty yleisiä arviointitekniikoiden toimintoja, missä punertavalla hehkulla on huomioitu valitut toiminnot.



Kuva 12. *Valitut toiminnot yleisistä arviointitekniikoiden toiminnoista.*

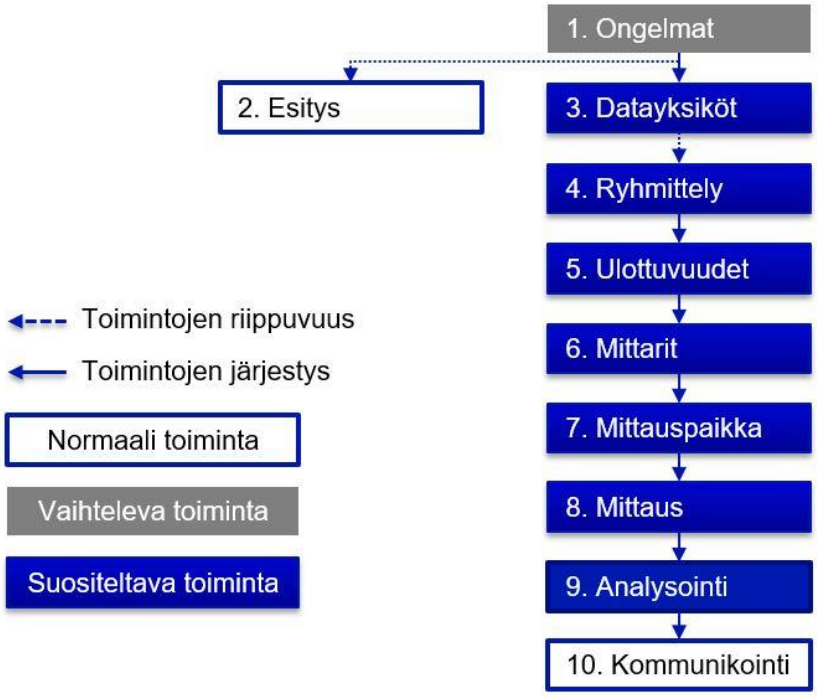
Valitut toiminnot ovat:

1. Organisaation ongelmien tunnistaminen ja priorisointi (ongelmat)
2. Arviointiprojektin esittäminen ylimmälle johdolle (esitys)
3. Datan ja attribuuttien valitseminen (datayksiköt)
4. Datan ryhmittely (ryhmittely)
5. Ulottuvuuksien tunnistaminen (ulottuvuudet)
6. Mittareiden tunnistaminen (mittarit)
7. Mittauspaikan valitseminen (mittauspaikka)
8. Objektiivisten ja subjektiivisten mittauksien suorittaminen (mittaus)
9. Tulosten analysointi (analysointi)
10. Tulosten kommunikointi ja jakaminen (kommunikointi)

Valittuihin toimintoihin on sisällytetty kaikki muut suositellut toiminnot, paitsi referenssidatan tunnistaminen. (Woodall et al. 2013) Se jätettiin pois, koska valitut objektiiviset mittarit eivät vaadi sen hyödyntämistä.

5.5 Arviointitoimintojen konfigurointi

Toimintojen konfiguroinnin tarkoituksena on järjestää toiminnot järkevään järjestykseen ja sisällyttää toimintojen riippuvuussuhteet. Kuvassa 13 on esitetty valittujen toimintojen järjestys ja suhteet.



Kuva 13. *Valitut toiminnot järjestyksessä.*

Ensimmäinen toiminto on kohdeyrityksen laatuongelmien tunnistaminen ja priorisointi. Laatuongelmia on jo aiemmin tunnistettu työtehtävien myötä, joten nyt päädyttiin vain listaamaan keskeisimmät ongelmat. Tunnistetut datahaasteet liittyvät erimuotoiseen dataan, virheellisiin merkkeihin, standardeista ja liiketoimintasäännöistä poikkeavaan dataan sekä puutteelliseen ja vanhaan dataan. Esimerkiksi nimissä voi olla lyhenteitä, sähköposteista saattaa puuttua loppupääte, datan arvoja puuttuu tai on merkitty tyhjään viittaavalla arvolla ja tekstimuotoisessa kentässä voi olla numeroita. Lisäksi asiakastietojen päivitystiheys vaihtelee, minkä vuoksi osa datasta voi olla vanhaa.

Toinen toiminto on arviointiprojektin esittäminen, minkä tarkoituksena on saada ylimmän johdon tuki projektille. Tässä tapauksessa arviointiprojektista toteutettiin esitys työn ohjaajille ennen empirian aloittamista. Sen keskiössä oli Hybridi-arviointimenetelmän eri vaiheiden läpikäyminen, painottuen tavoitteita ja vaatimuksia vastaavien arviointitoimintojen valitsemiseen.

Kolmas toiminto on datan ja attribuuttien valitseminen. Arvioitavaksi dataksi valittiin Master asiakasdata, joka on rakenteellista dataa. Attribuuteiksi valittiin asiakkaiden yhteystiedot, joihin kuuluvat nimi, henkilötunnus, puhelinnumero, sähköposti ja katuosoite.

Neljäs toiminto on datan ryhmittely, jossa datayksiköitä ryhmitellään eri luokkiin. Subjektiivisia mittauksia varten dataa ei ryhmitelty tarkemmin, vaan Master asiakasdatan yhteystietoja käsiteltiin kokonaisuutena. Objektiivisia mittauksia varten puolestaan dataa ryhmiteltiin, jotta datamassaa saatiin supistettua datan käsittelyn helpottamiseksi. Datan

keräämiskriteeriksi valittiin tietyn maakunnan keskustaajaman asiakkaat. Datan keräämisessä huomioitiin, että valitulla alueella ei korostu mikään tietty laatuongelma muuhun dataan verrattuna.

Viides toiminto on ulottuvuuksien tunnistaminen, jossa valitaan asianmukaisimmat ulottuvuudet laatuongelmat ja tavoitteet huomioiden. Objektiivisten mittareiden ulottuvuuksiksi valittiin ajantasaisuus, täydellisyys, virheettömyys ja sähköpostin osalta myös oikeellisuus. Objektiivisten mittareiden ulottuvuuksien valinnassa huomioitiin laatuongelmien ja tavoitteiden lisäksi myös mahdollisuus määrälliseen objektiiviseen mittaamiseen. Subjektiivisten mittauksien ulottuvuudet valittiin AIMQ-kyselylomakkeen perusteella, ja valitut ulottuvuudet ovat esitetty liitteessä A. Subjektiivisten mittauksien osalta haluttiin laajempaa näkemystä datan laadun tilasta, minkä vuoksi ulottuvuuksia valittiin yhteensä 15. Valituissa ulottuvuuksissa on pieniä eroavaisuuksia AIMQ-kyselylomakkeen ulottuvuuksien kanssa, sillä esimerkiksi ajantasaisuus erotettiin omaksi ulottuvuudeksi ja tulkittavuus sekä ymmärrettävyys yhdistettiin yhdeksi ulottuvuudeksi samankaltaisten ominaispiirteiden vuoksi.

Kuudes toiminto on datan laadun mittareiden tunnistaminen, jossa tunnistetaan, kehitetään tai käytetään olemassa olevia mittareita. Kaikkien valittujen ulottuvuuksien mittarit perustuvat kaavaan (1).

$$Suhde = 1 - \left[\frac{Epämieluisten\ tuloksien\ lukumäärä}{Kokonaislukumäärä}\right] \quad (1)$$

Seitsemäs toiminto on mittauspaikan valitseminen, jossa määritetään myös subjektiivisten mielipiteiden antajat eli haastateltavat henkilöt. Arvioitavaksi dataksi valittiin Master asiakasdata, joten mittauspaikkana toimii Master datan tietojärjestelmä, josta data siirretään Exceliin objektiivisia mittaustoimintoja varten. Haastateltavien valinnassa hyödynnettiin harkinnanvaraista otosta, sillä haastateltaviksi valittiin Master asiakasdatan käyttäjiä kolmesta ei työntekijäryhmästä. Työssä on painotettu käyttäjien näkökulman tärkeyttä laadun määrittämisessä, minkä vuoksi haastateltaviksi valittiin ainoastaan käyttäjiä. Haastatteluissa Master dataa ei rajattu tiettyyn tietojärjestelmään, sillä työntekijät käyttävät myös alkuperäisestä Master datasta luotuja näkymiä. Haastateltavat henkilöt ovat esitetty taulukossa 9.

Taulukko 9. *Haastateltavat työntekijät ja niiden lukumäärä.*

Haastateltavan rooli	**Lukumäärä**
Datatieteilijä	4
Kohderyhmäsuunnittelija	3
Suositteluautomaation työntekijä	3
Yhteensä	**10**

Kahdeksas toiminto on objektiivisten ja subjektiivisten mittauksien suorittaminen. Subjektiivisten mittauksien osalta hyödynnettiin haastattelua AIMQ-kyselylomakkeeseen pohjautuen, missä haastateltavat antoivat numeerisen arvion ja sanallisen perustelun laadun ulottuvuuksista. Numeeriset arvioinnit annetiin väliltä 0 (täysin eri mieltä) - 10 (täysin samaa mieltä). Haastatteluista saatiin siis laadullista ja määrällistä aineistoa. Haastattelut toteutettiin Skypen avulla ja ne nauhoitettiin niiden pätevyyden kasvattamiseksi ja laadullisen analyysin helpottamiseksi. Haastattelut olivat puolistrukturoituja, missä hyödynnettiin samaa haastattelurunkoa, mutta tarvittaessa esitettiin myös lisäkysymyksiä vastauksien tarkentamiseksi.

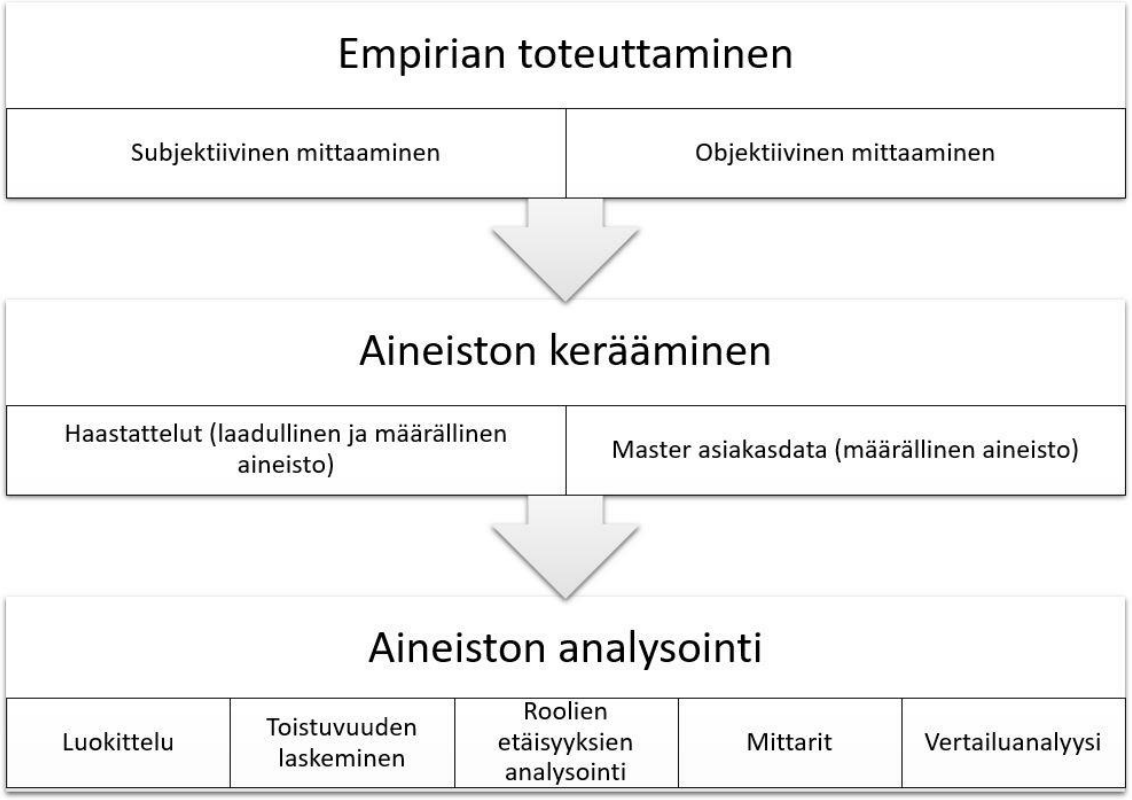
Objektiivisessa mittaamisessa hyödynnettiin kaavaa (1), jossa epämieluisten tuloksien lukumäärän sisältö riippui tarkasteltavasta ulottuvuudesta. Täydellisyyden osalta siihen sisällytettiin kaikki tyhjät ja tyhjään viittaavat arvot. Ajantasaisuuden kohdalla siihen valittiin kaikki ne asiakkaat, joiden viimeisimmästä tietojen päivityksestä oli yli 2 vuotta aikaa datan keräämisajankohtaan nähden. Virheettömyydessä hyödynnettiin kohdeyrityksen liiketoimintasääntöjä, joiden avulla saatiin selville muodoltaan ja sisällöltään validit arvot. Tämän myötä voidaan todeta, että virheettömyyttä tarkasteltiin validiuden näkökulmasta. Virheellisiksi eli epämieluisiksi tuloksiksi valittiin ne arvot, jotka eivät täyttäneet tarkasteltavan attribuutin liiketoimintasääntöjä. Sähköpostin oikeellisuuden kohdalla epämieluisia tuloksia olivat vahvistamattomat sähköpostit.

Yhdeksäs toiminto on tulosten analysointi. Haastatteluiden analysointien valmistelu aloitettiin litteroimalla jokaisen haastateltavan vastaukset omalle Excel-sarakkeelle ulottuvuuksittain. Litterointia ei toteutettu sanasta sanaan, vaan haastatteluista poimittiin olennaiset aiheeseen liittyvät asiat. Tämän jälkeen vastaukset ryhmiteltiin työntekijäryhmittäin, minkä jälkeen muodostettiin työntekijäryhmien koostetut vastaukset ulottuvuuksista. Tulosten yhteenvetoa varten vastauksia luokiteltiin niiden samankaltaisuuksien perusteella, minkä lisäksi huomioitiin niiden toistuvuus. Haastateltavien numeerisista vastauksista puolestaan laskettiin työntekijäryhmien sisäinen keskiarvo sekä työntekijäryhmien välinen keskiarvo, minkä jälkeen vastaukset visualisoitiin. Haastatteluiden eli subjektiivisten mittauksien yhtenä analysointimenetelmänä voidaan sanoa olevan myös roolien etäisyyksien analysointi, sillä visualisoinneista on nähtävillä työntekijäroolien vastauksien etäisyydet jokaisen ulottuvuuden kohdalla.

Objektiivisten mittauksien tuloksien analysoinnissa hyödynnettiin mittareiden arvoja. Ajantasaisuuden mittari on asiakasrivikohtainen, kun taas täydellisyyden, virheettömyyden ja oikeellisuuden mittarit ovat attribuuttikohtaisia. Jokaisen mittarin kohdalla esitetään itse mittarin ja tuloksien lisäksi myös mittausmenetelmä. Tulokset esitetään väliltä 0-1, jossa 1 viittaa korkeimpaan mahdolliseen laatuun ja 0 alhaisimpaan laatuun.

Viimeisenä analysointimenetelmänä hyödynnettiin DQA-arviointimenetelmässä esitettyä subjektiivisten ja objektiivisten mittauksien tuloksien vertailuanalyysiä. Siinä analysoi-

tiin ajantasaisuuden, täydellisyyden ja virheettömyyden tuloksia, koska kyseisiä ulottuvuuksia on tarkasteltu molemmissa objektiivisissa ja subjektiivisissa mittauksissa. Vertailuanalyysin tarkoituksena on nostaa esiin mahdollisia eroavaisuuksia subjektiivisten ja objektiivisten mittauksien tuloksien välillä. Kuvassa 14 on esitetty työssä käytetyt keskeiset empiiriset menetelmät.



Kuva 14. *Keskeiset empiiriset menetelmät.*

Viimeinen eli kymmenes valittu toiminto on tulosten kommunikointi ja jakaminen, jossa tulokset esitetään asiaankuuluville henkilöille. Tämä vaihe on käytännössä diplomityön tuloksien esittäminen kohdeyritykselle ja ohjaajille.

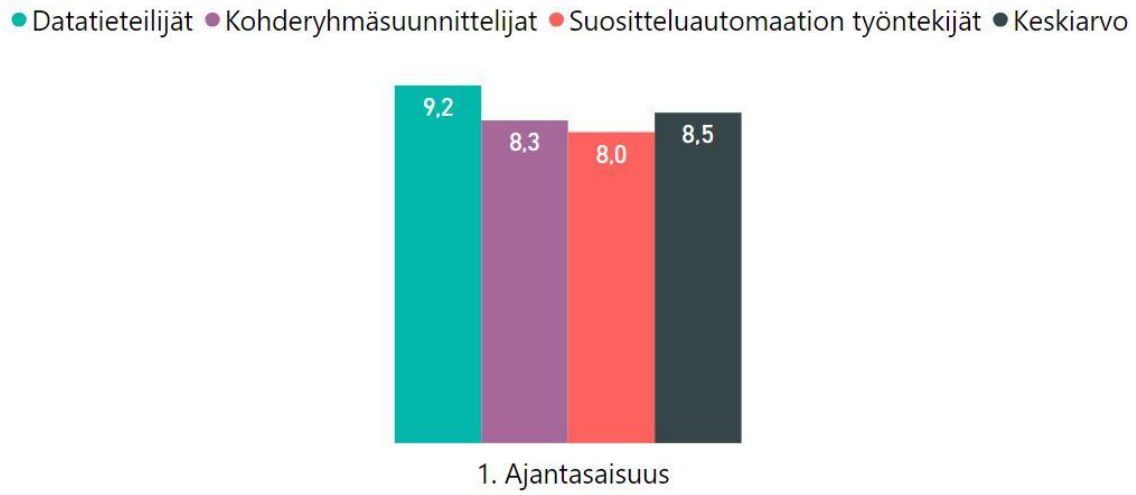
6. TULOKSET

Tutkimuksen tulokset on jaoteltu subjektiivisen ja objektiivisen mittaamisen tuloksien sekä näiden vertailuanalyysin tuloksien esittämiseen. Subjektiivisen mittaamisen tuloksissa tarkastellaan jokaista ulottuvuutta erikseen tuomalla esiin keskeiset asiat haastatteluista sekä visualisoinnit ulottuvuuksien numeerisista tuloksista työntekijäryhmien keskiarvon ja kokonaiskeskiarvon mukaan. Luvussa 6.2 esitetään yhteenveto haastatteluissa esiintyneistä haasteista ja kehitysehdotuksista toistuvuuden mukaan sekä yhteenvetokuvat ulottuvuuksien numeerisista tuloksista. Objektiivisen mittaamisen tuloksissa esitetään mittausmenetelmä, mittari ja tulokset. Lopuksi tarkastellaan subjektiivisen ja objektiivisen mittaamisen tuloksien eroavaisuuksia ajantasaisuuden, täydellisyyden ja virheettömyyden osalta.

6.1 Subjektiivisen mittaamisen tulokset

AJANTASAISUUS

Ajantasaisuus sai parhaimmat arvostelut datatieteilijöiltä, joille asiakkaiden yhteystiedot riittävät taustatiedoiksi. Markkinointitoimenpiteitä tekevät kohderyhmäsuunnittelijat ja suositteluautomaation työntekijät antoivat hieman alhaisemmat arvosanat. Eri työntekijäryhmien tuloksien keskiarvoksi saatiin 8,5. Kuvassa 15 on esitetty ajantasaisuuden numeeriset tulokset.



Kuva 15. *Ajantasaisuuden tulokset.*

Datatieteilijöiden mielestä Master asiakasdata on riittävän ajantasaista heidän tarpeisiin, eikä sen kohdalla ilmennyt puutteita. Asiakkaiden yhteystiedot riittävät heille taustatiedoiksi, sillä he eivät tee asiakaskontaktointia. Nykyinen järjestelmien viive riittää ehdot-

toman hyvin valtaosan tapauksissa, eikä tarvetta reaaliaikaiselle datalle tullut esiin. Suositteluautomaation työntekijöiden osalta puolestaan ilmeni tarve myös reaaliaikaiselle datalle. Esimerkiksi asiakkaan osoitteenmuutoksen tapahtuessa pitäisi käynnistyä ohjelma, jossa pohditaan asiakkaan tilannetta uudessa osoitteessa. Jos asiakas tekee tällä hetkellä muutoksen sunnuntain ja maanantain välisenä yönä, niin se on tiistaina työntekijöiden käytettävissä. Suositteluautomaation työntekijät totesivat kuitenkin myös, että kerran päivässä tapahtuva lataus riittää tällä hetkellä varsin hyvin.

”Meidän tarkastelemien asioiden sykli on yleensä viikko tai kuukausi, harvoin se on päiväkohtaista.” D3

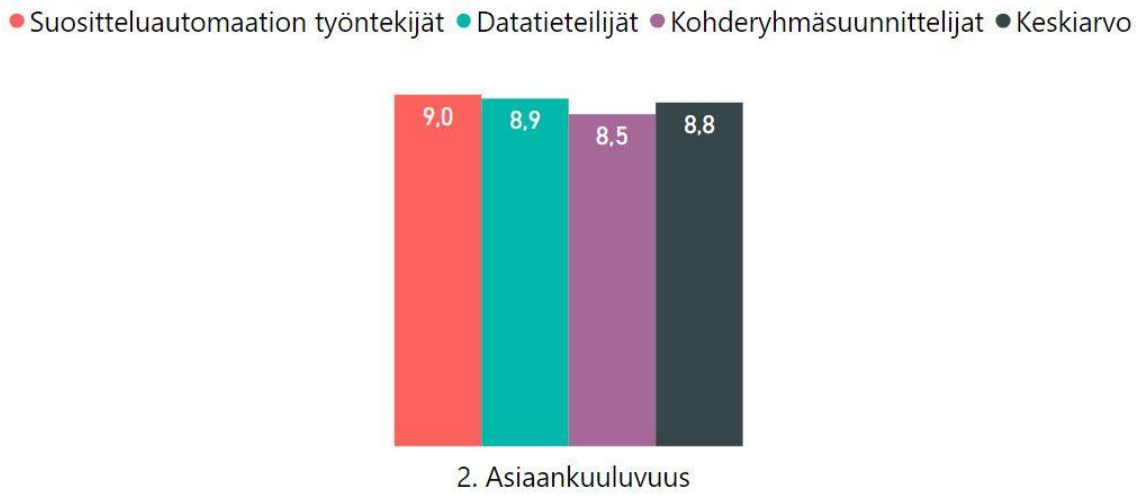
”Käyttäjän kannalta olisi ihanteellista, että data tulisi reaaliaikaisesti, vaikka se ei kuitenkaan ole realistista.” S1

Kohderyhmäsuunnittelijoiden ja datatieteilijöiden kohdalla ilmeni, että ajantasaisuudesta ei voi olla täysin varma, sillä puuttuu keinoja sen varmistamiseen. Lisäksi kohderyhmäsuunnittelijat totesivat, että yksittäisiä tapauksia voi olla, joissa data ei ole ajantasaista. Esimerkiksi markkinointilupien osalta on ilmennyt tapauksia, joissa on ollut haasteita ajantasaisuuden kanssa. Se ei kuitenkaan suoraan liity tarkastelun kohteena oleviin Master asiakasdatan yhteystietoihin. Pääasiallisiin markkinointitoimenpiteisiin datan koetaan olevan riittävän ajantasaista.

”Meidän työhön nähden datassa on aika paljon sellaista, mikä ei pidä paikkaansa. Se voi tosin olla ajantasaista.” K2

ASIAANKUULUVUUS

Asiaankuuluvuuden numeerisissa tuloksissa ei ilmennyt suuria eroavaisuuksia työntekijäryhmien välillä, ja keskiarvoksi saatiin 8,8. Kuvassa 16 on esitetty asiaankuuluvuuden numeeriset tulokset.



Kuva 16. *Asiaankuuluvuuden tulokset.*

Datatieteilijöiden näkökulmasta on erittäin hyödyllistä, kun ei tarvitse arpoa sen suhteen, mitä sähköpostia tai osoitetta käytetään. Datan päättelyprosessi on viety keskitetysti Master datan hallinnan tiimille ja samalla pois muiden työntekijöiden taakasta. Master asiakasdatassa oleva MDM-ID koettiin hyödylliseksi, koska se on huomattavasti tietoturvallisempi ja miellyttävämpi käyttää kuin henkilötunnus. Suositteluautomaation työntekijöiden työtehtävien kannalta Master asiakasdatan yhteystiedot ovat olennaisia, sillä puhelinnumeron ja sähköpostin avulla lähetetään asiakkaille viestejä. Datan koettiin soveltuvan työhön, sillä aiemmin se pohjautui asiakasnumeroihin, kun nyt se on henkilökeskeinen.

”Data on erittäin hyödyllistä. Työskentelen useiden eri datojen kanssa ja tämä on liimaa niiden välissä.” D2

”Kunhan me tunnistetaan se asiakas, se riittää meille.” D4

Datatieteilijät kokivat datan soveltuvuuden kohdalla haasteeksi sen, kun eri liiketoimintayksiköt eivät vielä hyödynnä MDM-ID:tä täysimääräisesti. Kohderyhmäsuunnittelijoiden kohdalla datan soveltuvuutta heikentää se, kun siihen on suodatettu dataa tiettyjen kriteerien perusteella. Lisäksi kohderyhmäsuunnittelijoiden ja datatieteilijöiden mielestä Master datassa olevien tuotetietojen luotettavuus on alhaista, minkä vuoksi niitä ei käytetä. Kohderyhmäsuunnittelijoiden kohdalla ilmeni myös, että joissain tapauksissa suoraan lähteestä saattaa saada myös parempaa dataa, jos haluaa esimerkiksi tietyn tuotteen asiakkuuden yhteystietoja. Ne eivät välttämättä ole samat, kun Master datassa olevat yhteystiedot.

”Oletusarvoisesti Master data on parempaa, kuin aiempi data. Se ei kuitenkaan tällä hetkellä vastaa täysin meidän tarpeitamme.” K2

Suositteluautomaation työntekijät totesivat, että tietyistä kanavista tulevilla asiakkailla ei välttämättä ole henkilötunnusta, ja osalla voi olla ainoastaan esimerkiksi sähköposti. Työtehtävien kannalta olisi parempi, jos myös näillä asiakkailla olisi samat tiedot kuin muista kanavista tulevilla asiakkailla. Kehitettävää olisi etenkin näiden tiettyjen kanavien rekisteröintiprosesseissa. Lisätarpeita ilmeni myös perheen yhteisten sopimusten tarjoamisessa, mutta toisaalta se ei välttämättä ole Master datan hallinnan tehtävä.

”Data on hyödyllistä ja todella tärkeää. Toisaalta esimerkiksi sähköposti olisi hyödyllisempi, jos siihen olisi yhdistettynä metadataa, kuten kannattaako siihen lähettää viestejä.” S2

ESITYKSEN JOHDONMUKAISUUS

Master asiakasdatan esityksen johdonmukaisuus koettiin olevan hyvällä tasolla, minkä myötä keskiarvoksi saatiin 9,2. Datan johdonmukaistamista voidaankin pitää yhtenä Master asiakasdatan keskeisimmistä tehtävistä. Kuvassa 17 on havainnollistettu esityksen johdonmukaisuuden tuloksia.

Kuva 17. *Esityksen johdonmukaisuuden tulokset.*

Datatieteilijät ovat havainneet pieniä epäjohdonmukaisuuksia, kun dataa tuodaan operatiivisiin järjestelmiin. On saattanut mennä sekaisin, mitkä ovat Master datan hallinnan tuottamia asiakasattribuutteja ja mitkä taas tuote- tai sopimusattribuutteja. Ulkomaalaisten osoitteiden kanssa voi olla myös epäjohdonmukaisuuksia, sillä esimerkiksi yhdysvaltalaisten postinumeroiden alussa voi olla osavaltiokoodi. Lisäksi henkilötunnuksissa on kirjaimia isoilla ja pienillä, mitkä pitäisi olla aina samassa muodossa. Jos datatieteilijät käyttävät henkilötunnuksia, niin ne muuttavat kaikki kirjaimet isoiksi.

”Data on hyvin johdonmukaista, kun asiakkaalla on lähtökohtaisesti esimerkiksi yksi kontaktinumero.” D1

Kohderyhmäsuunnittelijoiden mielestä data on varsin johdonmukaista, mutta pieniä poikkeamia voi löytyä. Dataa ei voi sanoa täysin johdonmukaiseksi, mutta se on kuitenkin hyvällä mallilla. Kohderyhmäsuunnittelijat olettavat ainakin niin, että esimerkiksi numeeriseksi määritellyssä kentässä on pelkästään numeroita. Ei ollut kuitenkaan varmuutta, onko kenttien määritykset tehty oikein. Suositteluautomaation työntekijät totesivat, että kenttien määrityksissä on ilmennyt epäjohdonmukaisuuksia, sillä esimerkiksi pelkästään numeroita sisältävä kenttä on määritelty tekstikentäksi. Syyksi epäiltiin sitä, että pelataan varman päälle. Se ei ainakaan kaadu siihen, että joku syöttää siihen kirjaimen.

”Datan nimitykset ja arvot ovat tietyssä muodossa.” K2

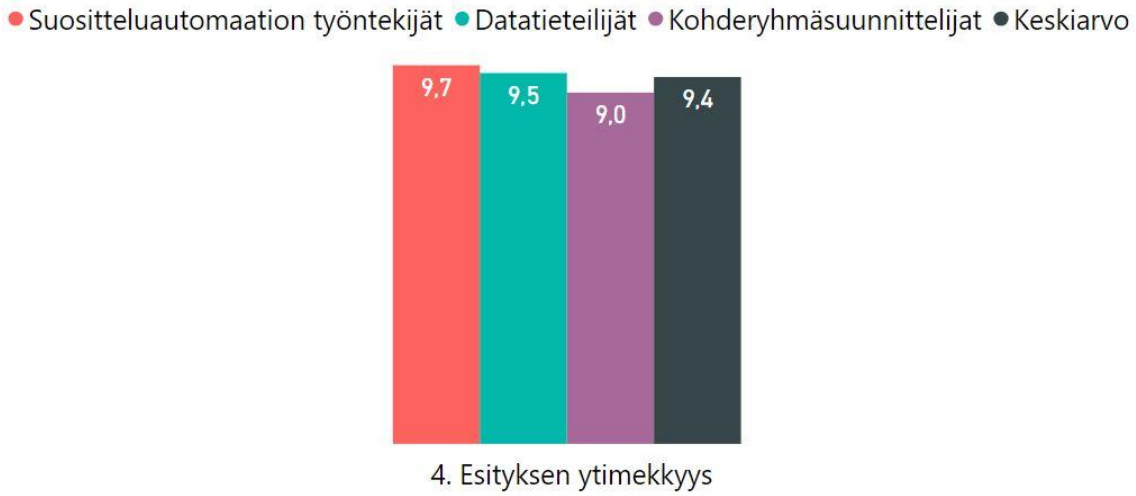
Suositteluautomaation työntekijöiden osalta ilmeni, että Master asiakasdatasta saattaa löytyä vielä asiakkaita, joilla on esimerkiksi etu- ja sukunimi väärin päin. Datassa on myös ruotsinkielisiä osoitteita, mikä ei täysin liity johdonmukaisuuteen. Käytön kannalta olisi kuitenkin hyödyllistä, jos kaikki osoitteet olisivat samalla kielellä. Datatieteilijöiden kohdalla ilmeni myös tarve hieman joustavammalle datalle, jotta asiakkaalla voisi olla

esimerkiksi useampia kontaktinumeroita. Se tosin saattaa mennä johdonmukaisuuden ulkopuolelle.

”Erittäin suuri osa datasta on esitetty johdonmukaisesti samassa muodossa.” S1

ESITYKSEN YTIMEKKYYS

Esityksen ytimekkyys sai ulottuvuuksista parhaimmat numeeriset tulokset, eikä ylimääräistä tai tarpeetonta data juurikaan havaittu datakentissä. Kuvassa 18 on esityksen johdonmukaisuuden tulokset, mistä nähdään eri työntekijäryhmien keskiarvon olevan 9,4.



Kuva 18. *Esityksen ytimekkyyden tulokset.*

Kaikki työntekijäryhmät kokivat Master asiakasdatan esitystavan tiiviiksi. Datatieteilijöiden mielestä data voisi olla vähemmänkin tiiviistä, sillä esimerkiksi osoitteista voisi erottaa kadunnimen ja -numeron toisistaan. Kohderyhmäsuunnittelijat eivät pidä mahdollista ylimääräistä dataa tarpeettomana, kunhan se on johdonmukaista.

”En ole havainnut, että taulurakenteissa tai muissa olisi niin sanotusti löysää.” D1

Kohderyhmäsuunnittelijoiden kohdalla ilmeni, että joissain kentissä voi olla tunnisteen lisäksi nimi. Sekin tosin riippuu käyttötarkoituksesta, onko se lopulta ylimääräistä dataa. Suositteluautomaation työntekijöiden osalta ei ilmennyt haasteita tai mahdollisia kehitystarpeita tämän ulottuvuuden kohdalla.

”Olen sitä mieltä, että yksi tieto per sarake on hyvä, ettei siihen kannata ketjuttaa muita tietoja.” K3

”En keksi, että tässä olisi mitään ongelmaa. Kyllä se on tiivistä dataa.” S2

HELPPOKÄYTTÖISYYS

Helppokäyttöisyyden osalta eri työntekijäryhmien numeeriset tulokset olivat linjassa keskenään. Kuvassa 19 on esitetty helppokäyttöisyyden tulokset, mistä nähdään eri työntekijäryhmien numeeristen vastausten keskiarvon olevan 9,0.

Kuva 19. *Helppokäyttöisyyden tulokset.*

Datatieteilijät eivät varsinaisesti muokkaa Master asiakasdataa. Sitä käytetään pitkälti niin, että haetaan MDM-ID ja siihen liittyviä asiakasattribuutteja, minkä jälkeen niitä yhdistetään muihin datoihin. Tosin pieniä muokkauksia saatetaan tehdä, kuten mahdollisten osavaltiokoodien poistaminen postinumeroiden edestä. Attribuuttien erilaiset nimeämiskäytännöt koettiin vaikeuttavan helppokäyttöisyyttä jonkin verran. Kohderyhmäsuunnittelijoiden mielestä dataa on helppo muokata vastaamaan tarpeita. Esimerkiksi kansainvälisessä muodossa olevia puhelinnumeroita pystytään vaivattomasti muuttamaan 0-alkuiseen muotoon. Tilanteen mukaan yhdistetään datoja myös samaan sarakkeeseen. Datan käsittely on helppoa, kun se on johdonmukaista. Suositteluautomaation työntekijöiden kohdalla datan muokkaamisen helppous koettiin melko vaikeaksi kysymykseksi, koska eri henkilöt tekevät muokkaukset tiettyjen speksien mukaan. Muokkaamista helpottaa se, kun katuosoite, postinumero ja postitoimipaikka ovat omissa kolumneissaan.

”Yrityksen tasolla ei ole selviä käytäntöjä attribuuttien nimeämisissä. Esimerkiksi puhelinnumeroa otsikoidaan eri datoissa erittäin monella eri tavalla.” D3

Datatieteilijöiden kohdalla ilmeni, että datan yhdistettävyyden osalta on ilmennyt haasteita operatiivisissa järjestelmissä. Eri liiketoiminnoilla ei välttämättä ole ollut aikaa ottaa MDM-ID:tä käyttöön, mikä aiheuttaa alipeittoa yhdistettävyydessä. Toiveena olisi, että MDM-ID:tä käytettäisiin enemmän ja työntekijät tietäisivät, mitä se tarkoittaa. Sen avulla Master asiakasdataa on helppo yhdistellä muun datan kanssa. Kohderyhmäsuunnittelijat totesivat myös, että datan yhdistettävyyttä vaikeuttaa se, kun Master datan avainarvoja ei välttämättä löydy muusta datasta. Suositteluautomaation työntekijät totesivat, että Master asiakasdatan kanssa ei ole ilmennyt ongelmia yhdistettävyyden kanssa, sillä sieltä löytyy

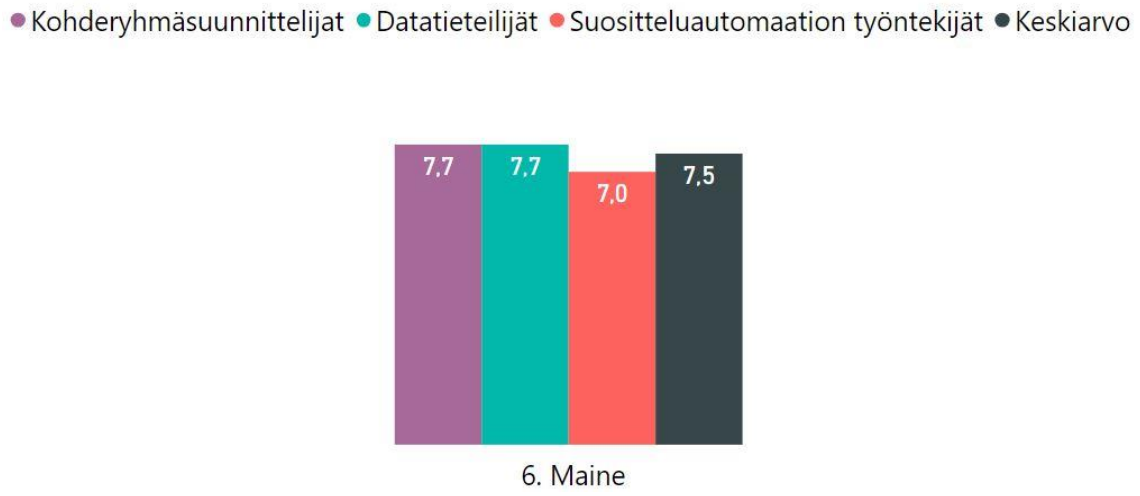
sekä henkilötunnus, että MDM-ID. Puutteita koetaan olevan enemmänkin muissa datoissa.

”On se helposti yhdistettävissä, kun ne on johdonmukaisesti nimetty. Ihan sarakkeen nimen perusteella pystyy yhdistelemään. Toki se on aina kiinni siitä muusta datasta.” K3

”Avainkentän (MDM-ID) käyttäminen on levinnyt jo melko laajalle yllättävänkin nopeasti.” S2

MAINE

Kaikki työntekijäryhmät olivat sitä mieltä, että Master datassa on muutamia haasteita, mitkä heikentävät sen mainetta. Numeeristen vastausten keskiarvoksi saatiin 7,5. Kuvassa 20 on esitetty maineen tulokset.



Kuva 20. *Maineen tulokset.*

Datatieteilijät totesivat Master asiakasdatan ytimellä olevan ihan hyvä maine. Kaikkeen Master dataan ei kuitenkaan pystytä tällä hetkellä luottamaan, mikä painottui arvosteluissa. Keskeisimpiin haasteisiin ja arvolupauksiin Master datan hallinta on vastannut, kuten asiakkaiden tunnistamiseen ja suostumusten hallintaan. Sen olisi kuitenkin voinut toteuttaa myös paremmin tietyissä kulmatapauksissa.

”MDM-projektin alkuvaiheissa puhuttiin, että MDM pelastaa, minkä myötä ladattiin kohtuullisen kovat odotukset.” D1

Datatieteilijät totesivat Master datassa ilmenneen useita ongelmia, sillä se on ollut välillä kaatuneena tai tyhjänä. Kohderyhmäsuunnittelijat kokivat maineen olevan hieman koetuksella markkinointilupiin liittyvien haasteiden vuoksi. Alkuperäisestä Master asiakasdatasta luodussa näkymässä on ollut myös haasteita esimerkiksi tuotetietojen kohdalla, sillä ne eivät ole pitänyt paikkaansa. Suositteluautomaation työntekijät totesivat myös,

että tuotetietojen kohdalla on ollut haasteita. Master dataa pidetään kuitenkin yrityksen parhaimpana datana, jolla on hyvä maine siltä osin, kun se toimii.

Kohderyhmäsuunnittelijat eivät pitäneet Master datan liiketoimintalogiikkaa täysin aukottomana. Nyt yritetään yhdistää esimerkiksi asiakkaan kaikki viisi asiakasnumeroa yhdeksi riviksi, minkä myötä on ilmennyt poikkeamia. Oletuksena valitut arvot ovat parhaimpia mahdollisia, mutta eivät ne kuitenkaan aina ole. Datatieteilijöiden mielestä suunnitteluvaiheessa tehty ratkaisu, että asiakkaaseen liittyy esimerkiksi yksi sähköposti ja yksi puhelinnumero, vaikutti hieman tarpeettomalta yksinkertaistukselta.

”Ne lupaukset, miten se tulee muuttamaan kaiken ja kaikki tulee toimimaan heti, niin se ei ole meidän työn kannalta mennyt ihan putkeen.” K2

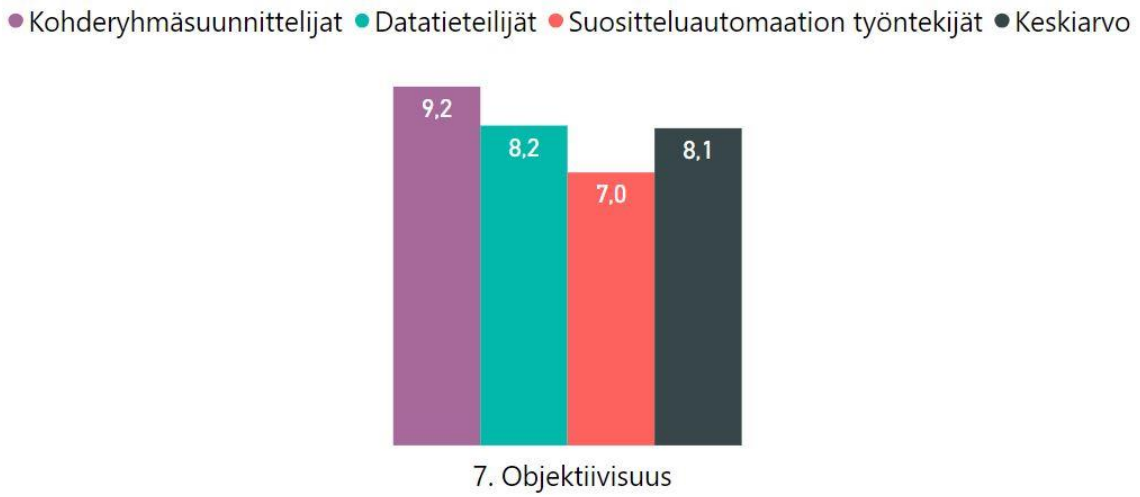
Suositteluautomaation työntekijät pitivät isona dataongelmana sitä, kun kaikkia sähköposteja ei vahvisteta asiakkailta. Ne tulisi vahvistaa aina, jotta asiakas on antanut oman oikean sähköpostinsa, eikä esimerkiksi kaverinsa tai väärin kirjoitettua sähköpostia. Suositteluautomaation työntekijöillä on myös oma logiikka toimimattomien sähköpostien suodattamiseen, jottei niihin kohdenneta markkinointia.

Datatieteilijät eivät osanneet vastata datalähteiden maineeseen, koska niitä ei tiedetty tai tunnistettu taustalla tapahtuvaa prosessia. Suositteluautomaation työntekijät ovat havainneet haasteita datalähteissä, sillä lähdejärjestelmässä data on täysin tietoisesti epäluotettavaa. Etenkin niitä lähteitä tulisi parantaa, missä ei vaadita tunnistautumista. Toisaalta ilmeni myös, että lähteitä on kehitetty paljon viime aikoina.

”Data ei tule hyvistä lähteistä, datalähteet ovat surkeita.” S2

OBJEKTIIVISUUS

Objektiivisuuden numeerisissa tuloksissa oli suuria eroja. Kohderyhmäsuunnittelijat kokivat Master asiakasdatan hyvinkin objektiiviseksi (9,2), kun taas suositteluautomaation työntekijät antoivat heikoimmat arvostelut (7,0). Toisaalta monet haastateltavat eivät tunteneet objektiivisuuteen vaikuttavaa datan keräysprosessia, minkä vuoksi kaksi haastateltavaa jätti vastaamatta tämän ulottuvuuden kysymyksiin. Objektiivisuuden tulokset on esitetty kuvassa 21.



Kuva 21. *Objektiivisuuden tulokset.*

Datatieteilijät totesivat datan olevan luultavasti objektiivista, eikä siitä oltu kovinkaan varmoja. Toisaalta siihen vaikutti se, kun datan keräämistapaa ei tiedetty tarkalleen. Suositteluautomaation työntekijät kokivat myös objektiivisuuden kysymykset melko vaikeiksi, koska ei ollut konkreettista tietoa Master asiakasdatan keräysprosesseista.

”Näyttää objektiivisesti kerätyltä, uskon sen olevan objektiivisesti kerättyä, mutta en voi olla siitä täysin varma.” S1

Datatieteilijöiden mielestä suostumusten hallinnassa on hyödynnetty inhimillistä perspektiiviä liiketoimintaohjauksen johdosta, minkä vuoksi se oltaisiin voitu toteuttaa hieman objektiivisemminkin. Lisäksi tuotetietojen haasteet nousivat esiin myös tässä ulottuvuudessa. Kaiken ei koettu olevan kunnossa, joten datan ei voitu sanoa perustuvan täysin faktoihin. Suositteluautomaation työntekijät eivät olleet täysin varmoja datan faktuaalisuudesta, mutta sen osalta nousi esiin myös kehitettävää. Esimerkiksi sähköpostien osalta asiakas voi antaa väärän sähköpostin vahingossa tai tarkoituksella, minkä todettiin olevan prosesseihin viittaava ongelma.

”Tuoteomistukset eivät tällä hetkellä mene kovin loogisesti. Esimerkiksi minulla ei ole lapsia, ja datan mukaan nuorimman lapsen ikä on 7-12 vuotta.” D3

Kohderyhmäsuunnittelijat kokivat Master asiakasdatan perustuvan pitkälti käyttäjien tarpeisiin, sillä ne datat on otettu käyttöön, mitä on tarvittu. Kaikkea dataa ei voi kuitenkaan heti tuoda, niin on todennäköisesti pitänyt priorisoida, että tuodaan ensin tärkeimmät attribuutit. Kaikkia asiakkaita ei tuoda Master dataan, mikä osaltaan vaikuttaa objektiivisuuteen. Esimerkiksi, jos asiakkaalta löytyy Y-tunnus, niin sitä ei oteta henkilöasiakkaiden Master asiakasdataan.

”Jos datassa ilmenee ongelmia, niin ne johtuvat muista syistä kuin objektiivisuuden puutteesta.” K1

OIKEA-AIKAISUUS

Kaikki kolme työntekijäryhmää kokivat Master asiakasdatan olevan tällä hetkellä riittävän oikea-aikaista, minkä myötä numeeristen vastausten keskiarvoksi saatiin 9,2. Kuvassa 22 on esitetty oikea-aikaisuuden tulokset.

Kuva 22. *Oikea-aikaisuuden tulokset.*

Datatieteilijät totesivat datan olevan käytettävissä niin oikeaan aikaan, kun sen toivookin olevan. Kohderyhmäsuunnittelijoiden mielestä on vaikea kokea, ettei data olisi ollut käytettävissä oikeaan aikaan. Suositteluautomaation työntekijät eivät ole havainneet ongelmia oikea-aikaisuuden kohdalla. Lisäksi tietojärjestelmien todettiin olevan sellaisia, että niistä saa datan silloin, kun sitä tarvitsee.

”Ei ole esim. sellaista tilannetta, että meidän pitäisi odottaa kuun 15. päivään asti, jotta se olisi päivittynyt ja voitaisiin edetä.”D3

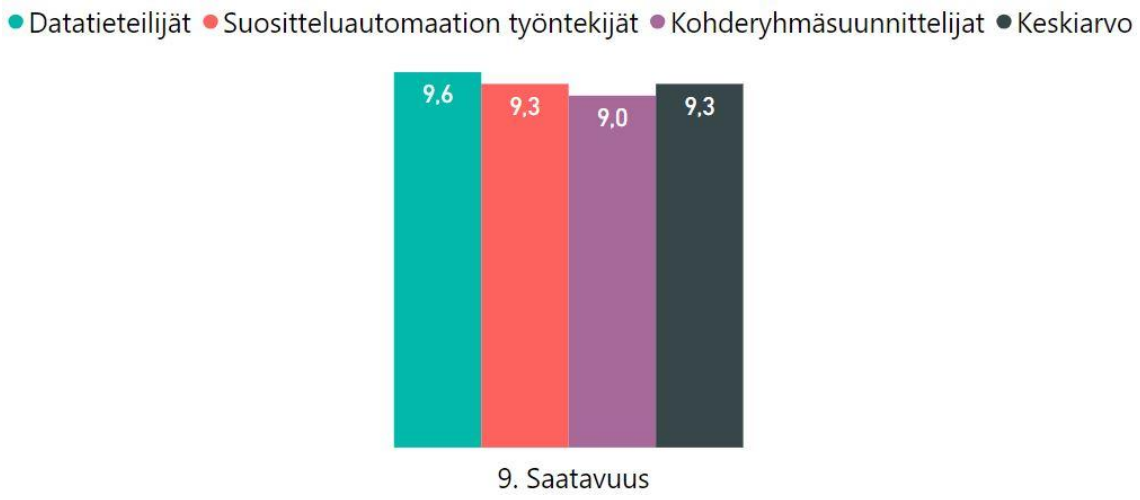
”Vuorokauden välein tapahtuva lataus on ihan riittävä, ainakin toistaiseksi.” K2

Kohderyhmäsuunnittelijoiden osalta ilmeni, että markkinointilupien kohdalla tietyt valinnat eivät ole aina välittyneet oikein. Joskus on myös saattanut käydä niin, että näkymä on ollut tyhjillään. Näin ei ole kuitenkaan käynyt haastateltavien kohdalla. Suositteluautomaation työntekijöiden kohdalla ilmeni myös mahdollinen tulevaisuuden tarve reaaliaikaiselle datalle. Esimerkiksi tulevaisuudessa saattaa olla sellainen liiketoimintatarve, jossa muuttoilmoituksesta halutaan tieto heti. Tämän myötä asiakkaita voitaisiin kontaktoida markkinointiautomaation keinoin riittävän nopeasti.

”Tällä hetkellä data on käytettävissä aina oikeaan aikaan meidän käyttötarkoituksia varten.” S1

SAATAVUUS

Saatavuuden numeeristen vastausten keskiarvoksi saatiin 9,3. Kuvassa 23 on esitetty saatavuuden tulokset.



Kuva 23. *Saatavuuden tulokset.*

Datatieteilijät kokivat Master asiakasdatan olevan helposti ja nopeasti saatavissa, kun siihen on pääsyoikeudet ja tietää sen sijainnin. Suositteluautomaation työntekijät kokivat myös datan saatavuuden olevan kunnossa, eikä heiltä ilmennyt haasteita tai lisätarpeita tämän ulottuvuuden kohdalla.

”Alkuperäinen Master data on helposti saatavissa ja modifioitu näkymä on helposti saatavissa.” S1

Datatieteilijöille alkuperäinen Master asiakasdata ei ole saatavilla, mutta sitä ei pidetty ongelmana. Operatiivisissa järjestelmissä datalle saatetaan tehdä muutoksia tiettyihin tarpeisiin, minkä myötä se saattaa olla saatavissa hieman eri muodoissa eri paikoissa. Kohderyhmäsuunnittelijat totesivat Master asiakasdatan tietokantataulujen isot koot hidastavaksi tekijäksi saatavuuden nopeuteen. Sen vuoksi muutoksien prosessointi vaatii aikaa, mikä tosin riippuu myös työvälineestä. Todettiin myös, että ei ole hyviä työkaluja päästä alkuperäiseen Master asiakasdataan kiinni. Lisätarpeena esitettiin mahdollisuus pystyä tekemään itse näkymät, joita voisi hyödyntää yleisesti käytettävillä työkaluilla.

”Data on helposti saatavissa. Haasteena on ehkä se, että se on saatavissa vähän turhankin monessa paikassa.” D1

”Dataan pääsee helposti käsiksi ja tarvittavan nopeasti.” K2

SOPIVA MÄÄRÄ

Sopivaa määrää tarkasteltiin pääosin rivimäärien näkökulmasta, koska tarkasteltavat attribuutit rajattiin yhteystietoihin. Kuvassa 24 on esitetty sopivan määrän tulokset, mistä nähdään eri työntekijäryhmien numeeristen vastausten keskiarvon olevan 8,9.

Kuva 24. *Sopivan määrän tulokset.*

Datatieteilijät ja suositteluautomaation työntekijät kokivat Master asiakasdatan rivimäärän yleisesti sopivaksi. Toisaalta datatieteilijät toivoivat tiettyjen liiketoimintojen asiakkaiden osalta laajempaa MDM-ID:n hyödyntämistä. Ei oltu myöskään täysin varmoja, säilytetäänkö vanhojen asiakkaiden tietoja. Esimerkiksi, kuka saattaisi tulla takaisin yrityksen asiakkaaksi. Suositteluautomaation työntekijät totesivat Master datassa olevan lähes yhtä paljon asiakkaita, kuin yrityksellä on asiakkaita. Datan määrän koettiin myös avaavan yritykselle sellaisia mahdollisuuksia, mitä se ei hyödynnä.

”Sieltä ei mitään selkeitä tuote- tai asiakaslatauksia puutu.” D3

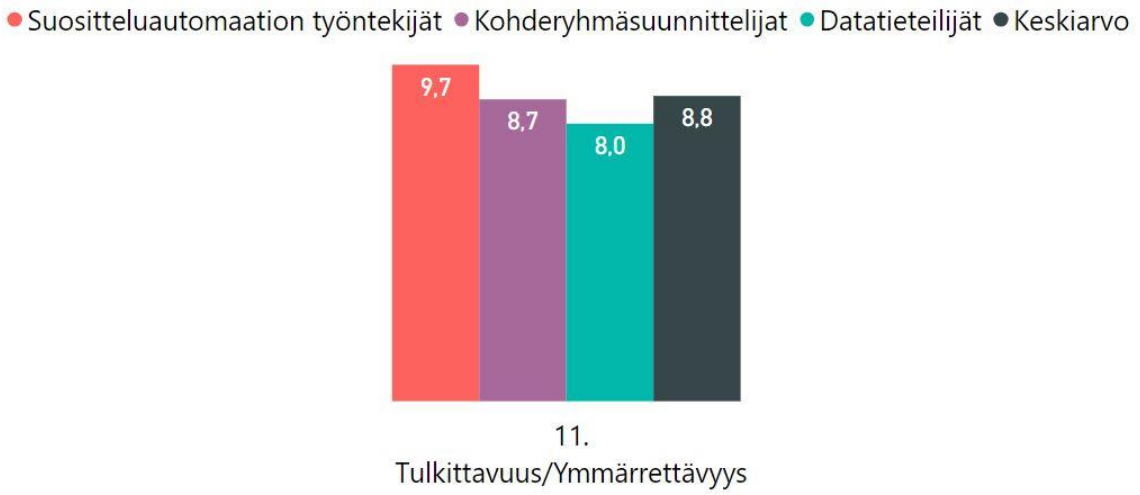
”Master datassa on ne, mitä siellä pitääkin olla. En näe siinä ongelmaa.” S2

Kohderyhmäsuunnittelijoiden kohdalla datan määrän ei todettu riittävän kaikkiin tarpeisiin, mutta useimpiin tarpeisiin se riittää hyvin. Tarve olisi yrityksen kaikkien asiakkaiden datalle, sillä siellä ei ole esimerkiksi yritysasiakkaita. Normaalisti kohderyhmän koko on väliltä 10 000 – 100 000, minkä pystyy yleensä Master asiakasdatasta tuottamaan. Jos taas haluttaisiin Master asiakasdatan kautta tiedottaa asiakkaita, niin sieltä ei löytyisi kaikkia.

”Dataa on siinä mielessä liian vähän, kun Master asiakasdataan on valittu asiakkaat tiettyjen arvojen ja kriteerien perusteella.” K1

TULKITTAVUUS/YMMÄRRETTÄVYYS

Suositteluautomaation työntekijät antoivat tulkittavuudelle ja ymmärrettävyydelle korkeimman tuloksen (9,7), kun taas datatieteilijät antoivat heikoimman (8,0). Tulkittavuuden ja ymmärrettävyyden tulokset on esitetty kuvassa 25.



Kuva 25. *Tulkittavuuden ja ymmärrettävyyden tulokset.*

Kaikki työntekijäryhmät kokivat asiakkaiden yhteystiedot helposti tulkittaviksi ja ymmärrettäviksi, eikä suositteluautomaation työntekijöiltä ilmentynyt haasteita tai lisätarpeita tämän ulottuvuuden kohdalla. Toisaalta datatieteilijöiden ja kohderyhmäsuunnittelijoiden mielestä siellä on yhteystietojen lisäksi myös sellaisia attribuutteja, joista ei välttämättä suoraan osata sanoa niiden merkitystä. Datatieteilijät kokivat Master asiakasdatan dokumentaation puutteelliseksi, sillä ei tiedetty, että löytyykö siitä julkista dokumenttia. Lisäksi kaivattiin dokumentaatiota datan arvojen päättelyprosessista, miten esimerkiksi sähköpostin arvo asiakkaalle valitaan. Master asiakasdatan taustalla tapahtuva prosessi oli siis hyvinkin tuntematon.

”Olisi parempi kysymys, jos tarkasteltava data olisi monimutkaisempi.” S1

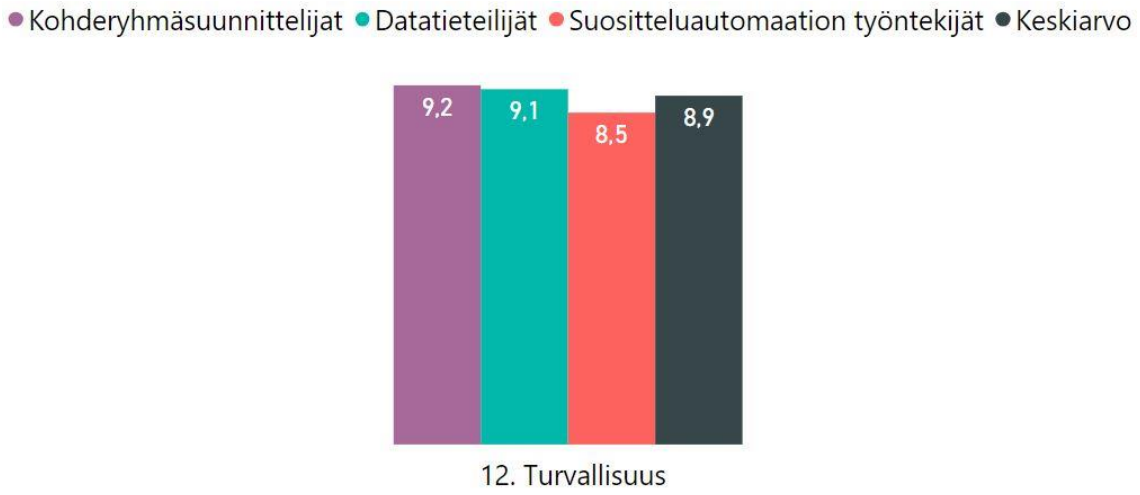
”Datan prosessointia ei olla dokumentoitu läheskään niin hyvin, kun se olisi pitänyt.” D1

Kohderyhmäsuunnittelijoiden kohdalla ilmeni, että yrityksellä on data standardeja, mutta niitä ei käytetä. Jos ei ole lukenut dokumentaatiota, niin esimerkiksi joidenkin aikaan liittyvien attribuuttien ymmärtäminen voi olla vaikeaa. Lisäksi Master datassa sopimusnumerot ja asiakasnumerot on päinvastoin nimetty muun datan kanssa. Edelleen käytetään myös vanhempien tietojärjestelmien datoja, joissa nimeämiset saattavat olla erilaiset.

”Asiakasnumero-nimikkeellä oleva data Master datassa onkin sopimusnumero toisessa datassa.”K2

TURVALLISUUS

Turvallisuuden numeeristen vastausten keskiarvoksi saatiin 8,9. Kuvassa 26 on esitetty turvallisuuden tulokset.



Kuva 26. *Turvallisuuden tulokset.*

Kaikki työntekijäryhmät uskoivat Master asiakasdatan olevan suojattu luvattomalta pääsyltä, sillä sen käyttämiseen vaaditaan oikeudet. Datatieteilijöiden kohdalla ilmeni oikeuksien lisäksi myös, että esimerkiksi SAS-työkalua käyttäviltä vaaditaan turvallisuusselvitys, mikä menee yrityksen prosessien mukaan. Toisaalta Master dataa käytetään eri operatiivisissa järjestelmissä, missä se saattaa olla Master datan omien tietoturvakäytäntöjen ulkopuolella olevien henkilöiden käytettävissä. Tällöin se on operatiivisen järjestelmän vastuulla, onko data saatavissa vain tarkoitetuille henkilöille.

”Alkuperäiseen Master asiakasdataan ei ole minullakaan pääsyä, joten sen suhteen on tehty hyvää työtä turvallisuuden eteen.” D1

Kohderyhmäsuunnittelijat kokivat, että turvallisuuden miettiminen ei ole heidän tehtävä. Haastateltavat tuntuivat pääsevän eri datoihin melko hyvin, mutta nousi esiin myös joitain paikkoja, mihin olisi kiva päästä nopeammin. Todettiin myös, että Master data ei välttämättä ole saavutettavissa kaikille niille, jotka sitä tarvitsisi.

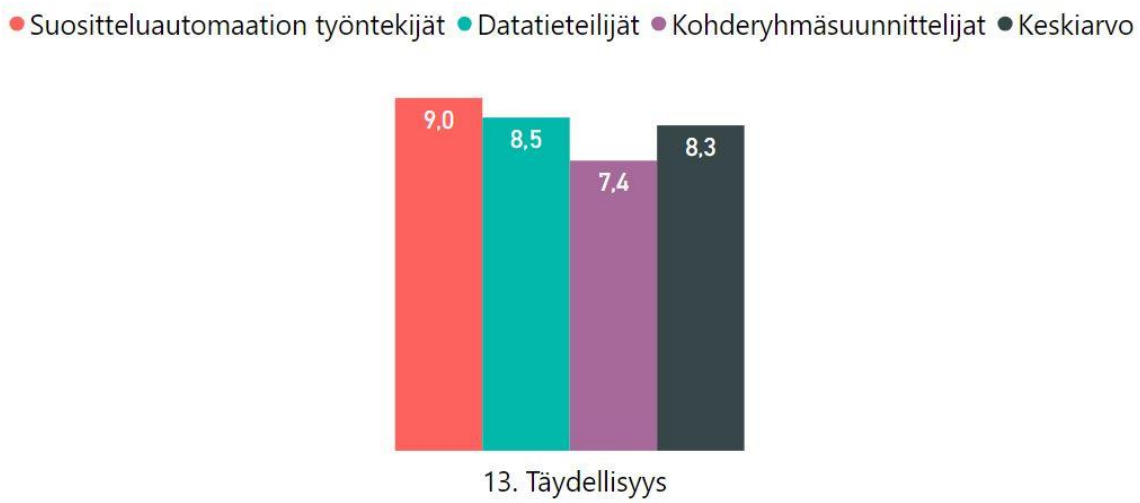
”Tunnukset vaaditaan ja kirjaudutaan, niin eiköhän se ole ihan ok-tasolla.” K3

Suositteluautomaation työntekijöiden kohdalla ilmeni, että dataa ei ole suojattu luvattomalta pääsyltä absoluuttisesti, mutta kuitenkin tarpeeksi suojattu. Jos on oikeudet Master datan johdannaiseen dataan, niin sitä pystyy jakamaan toimistoverkossa henkilöille, joilla ei ole siihen oikeuksia. Ihannetilanteessa olisi yksi admin, jonka kautta oikeudet annetaan. Kaikilla ei ollut oikeuksia alkuperäiseen Master asiakasdataan, minkä vuoksi suojauksen koettiin olevan riittävän hyvällä tasolla.

”Henkilöt on rajattu tarkkaan. Lukuoikeudet meillä on, kirjoitusoikeuksia emme tarvitse.” S3

TÄYDELLISYYS

Suositteluautomaation työntekijöiden vastausten keskiarvo oli korkein (9,0), kun taas kohderyhmäsuunnittelijoiden keskiarvo oli alhaisin (7,4). Täydellisyyden numeeristen vastausten keskiarvoksi saatiin 8,3. Kuvassa 27 on esitetty täydellisyyden numeeriset tulokset.



Kuva 27. *Täydellisyyden tulokset.*

Datatieteilijät kokivat Master asiakasdatan ytimen olevan riittävän täydellistä heidän tarpeisiin. Se nähdään erinomaisena pohjana ja kohtuullisen kattavana näkymänä yhteen asiakkaaseen. Datatieteilijät eivät olleet kohdanneet juurikaan tyhjiä arvoja, ainakaan henkilötunnuksen ja puhelinnumeron kohdalla. Dataa voisi olla heidän mielestään enemmänkin. Uutta dataa kontrolloidaan, mutta historiadatan vuoksi sieltä voi löytyä poikkeamia.

”Aikojen alussa ei ollut mitään kontrollia, mitä sinne pystyi syöttämään. Historiasta johtuen sieltä voi löytyä hämäriä juttuja.” D4

Kohderyhmäsuunnittelijat totesivat, että datalähteessä on tyhjiä ja tyhjään viittaavia arvoja, minkä vuoksi niitä on myös Master datassa. Nyt on käytännössä yksi lähde, mutta useiden lähteiden tapauksessa tyhjiä arvoja ei saisi olla niin paljoa. Kohderyhmäsuunnittelijat eivät pysty tällä hetkellä tekemään työtehtäviä pelkästään Master datalla, minkä vuoksi sen ei todettu olevan riittävän täydellistä. Tarve olisi uusille attribuuteille ja datoille, jotta Master dataan voisi siirtyä 100 prosenttisesti. Ilmeni myös tarve pilkotulle osoitteelle, jossa olisi erikseen esimerkiksi kadunnimi, numero ja rappu. Markkinointilupiin liittyen on kohdattu ongelmia, sillä ne ovat olleet esimerkiksi vanhoja tai jollain tavalla erilaista tietoa Master datassa kuin muualla.

”Master datassa yhteysnumero voi olla tyhjä, vaikka asiakkaalla on jollain asiakasnumerolla sille arvo. Nämä ovat aika monimutkaisia juttuja.”K3

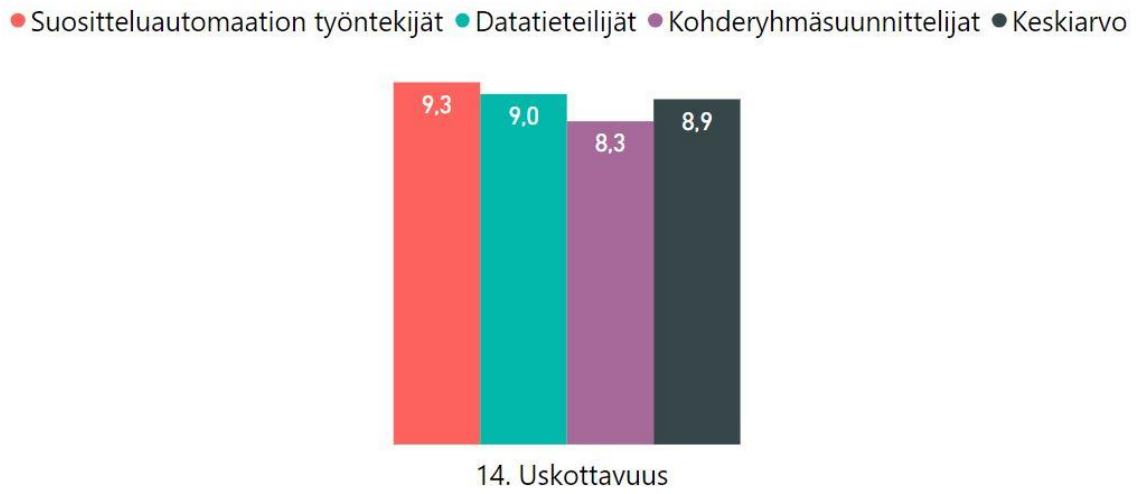
Suositteluautomaation työntekijät kokivat suuren osan datasta sisältävän kaikki tarvittavat arvot, etenkin perustietojen kohdalla. Todettiin, että kevyen tunnistautumisen lähteistä

tulee epätäydellistä dataa, minkä vuoksi etenkin alkuvaiheen rekisteröintiprosesseissa olisi kehitettävää. Lisäksi ilmeni myös, että tietoa on sekin, ettei ole tietoa.

”Tavallaan se ei ole Master datasta johtuva ongelma, jos sieltä puuttuu joku arvo. Enemmänkin yrityksen prosesseista johtuva ongelma, kun on suhtauduttu huonosti datan laatuun.” S2

USKOTTAVUUS

Uskottavuuden numeeriset tulokset on esitetty kuvassa 28, josta nähdään numeeristen vastausten keskiarvon olevan 8,9.



Kuva 28. *Uskottavuuden tulokset.*

Datatieteilijät pitivät Master asiakasdataa pääosion luotettavana ja lähtökohtaisesti uskottavana. Uskottavuuteen vaikuttaa kuitenkin datassa olevat testihenkilöt, jotka tulisi siivota sieltä pois.

”Osittain se on uskottavaa, osittain ei.” D3

Kohderyhmäsuunnittelijat kokivat Master asiakasdatan yhteystietojen olevan uskottavia ja luotettavia, mutta Master datan kokonaisuuteen liittyy muutamia uskottavuuteen liittyviä haasteita. Etenkin tuote- ja markkinointilupatietojen haasteet ovat vaikuttaneet uskottavuuteen. Toisaalta todettiin myös, että dataa ei voisi käyttää, jos sitä ei pitäisi uskottavana. Tietyt rajoitteet uskottavuuteen liittyen on hyvä tietää.

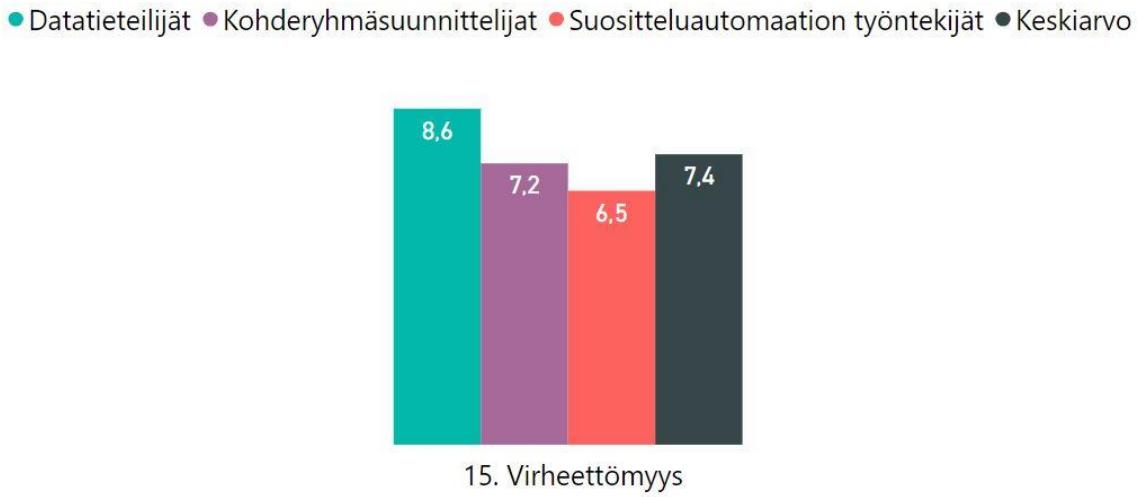
”En voi oikeastaan käyttää Master dataa täysin sokkona, vaan sen todenmukaisuus pitää vielä varmistaa.” K2

Suositteluautomaation työntekijät totesivat Master asiakasdatan olevan uskottavaa, sillä se ainakin näyttää uskottavalta. Toimimattomat sähköpostit koettiin ainoana uskottavuuteen liittyvänä haasteena.

”Sähköpostia lukuun ottamatta se on uskottavaa.” S3

VIRHEETTÖMYYS

Kaikki työntekijäryhmät olivat sitä mieltä, että data ei ole virheetöntä. Kuvassa 29 on esitetty virheettömyyden numeeriset tulokset, mistä nähdään keskiarvon olevan 7,4.



Kuva 29. *Virheettömyyden tulokset.*

Datatieteilijät totesivat, että datassa on pieniä poikkeamia. Virheellisyyksiä on kohdattu esimerkiksi tuotetietojen kohdalla, ja datassa olevien testihenkilöiden koettiin heikentävän tarkkuutta. Yksi attribuutin arvo per asiakas saattaa olla myös riittämätön tietyissä tapauksissa, mikä ei välttämättä liity suoraan datan tarkkuuteen.

”Ei se tietenkään virheetöntä ole, pieniä poikkeamia sieltä löytyy. Tavallaan olen oppinut hyväksymään, mitä se on ja tulen sen kanssa toimeen.” D4

Kohderyhmäsuunnittelijoiden osalta ilmeni, että virheellisyyteen liittyvät etenkin datan alkuvaiheen prosessit, jotka saattavat tuottaa virheellistä dataa. Data ei ole virheetöntä lähteessäkään, minkä vuoksi virheitä esiintyy myös Master datassa. Ainakin jossain vaiheessa oli tapauksia, joissa asiakkuudelta poistettu sähköposti ei poistunut Master datasta. Tämän vuoksi markkinointiviestejä on lähetelty sellaisiin paikkoihin, mihin ei välttämättä olisi saanut. Markkinointilupa-asiat koettiin myös yhdeksi keskeiseksi tekijäksi, mikä vaikuttaa virheellisyyteen.

”Tuntuu siltä, että automatiikkaa puuttuu virheellisyyksien osoittamiseen ja myös resursseja virheellisyyksien korjaamiseen.” K1

Kohderyhmäsuunnittelijoiden osalta ilmeni myös, että joistain kanavista voi tulla myös nimi ja henkilötunnus -pareja, jotka eivät täsmää. Nousi myös esiin, että olisi hyvä saada

tieto siitä, jos asiakas on tunnistautunut jollain yrityksen hyväksymällä tavalla, kuten TUPAS-tunnistautumisella. Standardista poikkeavia arvoja löytyy, mutta kokonaisuutena Master asiakasdatan koettiin toimivan hyvin pääasiallisiin markkinointitarpeisiin.

”Jos me käsitellään esimerkiksi miljoona riviä kahden viikon aikajanalla, ja sieltä tulee vaikka 10 tapausta takaisin, niin prosentuaalisesti se on tosi hyvin.” K2

Suositteluautomaation työntekijät kokivat Master asiakasdatan osittain virheelliseksi. Tietyistä kanavista tulevilla asiakkailla voi olla esimerkiksi sähköposteja, jotka eivät ole käytössä. Sähköposteissa on tunnistettu selkeitä, ilmeisesti suhteellisen helpostikin korjattavia virheitä, kuten vääriä loppupäätteitä ja kirjoitusvirheitä. Data on usein tarkkaa niiden asiakkaiden osalta, joilta on mahdollisuus saada kaikki tiedot. Todettiin myös, että datassa ei ole merkittävää isoa virhettä, mikä johtuisi Master datasta.

”Datan laatua voi parantaa, mikä on enemmänkin datalähteiden ja Master datan logiikan parantamista.” S2

6.2 Subjektiivisen mittaamisen tuloksien yhteenveto

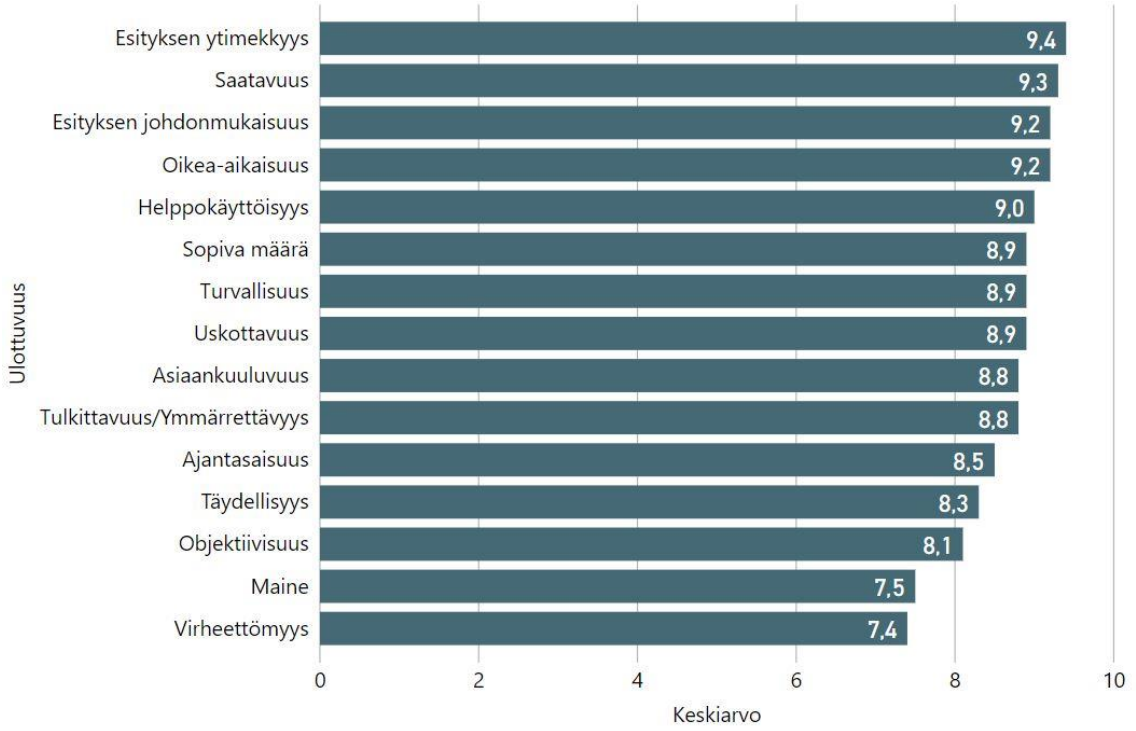
Haastatteluiden numeeristen vastausten eroavaisuuksien selkeyttämiseksi muodostetiin visualisointi työntekijäryhmien sisäisen keskiarvon avulla. Eroavaisuuksia korostettiin hyödyntämällä eri värejä, joiden vaihteluväli on 0,5. Punainen väri edustaa alhaisinta arvoa ja tumman vihreällä kuvataan korkeimpia arvoja. Kuvassa 30 on esitetty työntekijäryhmien numeeriset vastaukset värien mukaan jaoteltuna.

	Datatieteilijät	Kohderyhmäsuun.	Suositteluautomaat.
Ajantasaisuus	9.20	8.30	8
Asiaankuuluvuus	8.90	8.50	9
Esityksen johdonmukais.	9	9.70	9
Esityksen ytimekkyys	9.50	9	9.70
Helppokäyttöisyys	8.90	9	9.20
Maine	7.70	7.70	7
Objektiivisuus	8.20	9.20	7
Oikea-aikaisuus	10	9	8.70
Saatavuus	9.60	9	9.30
Sopiva määrä	9.20	8.70	8.70
Tulkittavuus/Ymmärret.	8	8.70	9.70
Turvallisuus	9.10	9.20	8.50
Täydellisyys	8.50	7.40	9
Uskottavuus	9	8.30	9.30
Virheettömyys	8.60	7.20	6.50

Kuva 30. *Haastateltavien numeeriset vastaukset värien mukaan jaoteltuna.*

Kuvasta 30 on nähtävissä, että kaikista alhaisimman numeerisen arvon (6,50) antoivat suositteluautomaation työntekijät virheettömyydelle, mikä on merkitty kuvassa punaisella. Korkeimman numeerisen arvon (10) puolestaan antoivat datatieteilijät oikea-aikaisuudelle, mikä on merkitty kuvassa tumman vihreällä. Suurimpia vaihteluita eri työntekijäryhmien keskiarvojen välillä oli objektiivisuuden, täydellisyyden ja virheettömyyden kohdalla.

Lisäksi muodostettiin yhteenvetokuva kaikkien haastateltavien numeeristen vastausten keskiarvosta ulottuvuuksittain. Kuvassa 31 on esitetty subjektiivisen mittaamisen tulokset haastateltavien kokonaiskeskiarvon mukaan.



Kuva 31. *Subjektiivisen mittaamisen tulokset.*

Tuloksista nähdään, että korkeimmat numeeriset arvot saivat esityksen ytimekkyys (9,4), saatavuus (9,3), esityksen johdonmukaisuus (9,2) ja oikea-aikaisuus (9,2). Alhaisimmat numeeriset arvot saivat puolestaan virheettömyys (7,4), maine (7,5), objektiivisuus (8,1) ja täydellisyys (8,3). Kokonaisuutena ulottuvuuksien numeeriset arvot painottuivat yllättävänkin korkealle tasolle, ottaen huomioon numeerisen skaalan 0-10.

Taulukossa 10 on esitetty keskeisimmät haastatteluista ilmenneet haasteet ja kehitysehdotukset. Kyseiseen taulukkoon on otettu mukaan kaikki ne asiat, jotka esiintyivät haastatteluissa vähintään kaksi kertaa. Haasteita ja kehitysehdotuksia ei olla luokiteltu sanasta sanaan, vaan samaan haasteeseen tai kehitysehdotukseen on sisällytetty siihen viittaavat asiat. Huomioitavaa on, että samalta haastateltavalta on voinut ilmentyä sama asia eri ulottuvuuksissa. Jos taas sama asia on ilmennyt useasti tietyssä ulottuvuudessa, niin tällöin sen on nostanut esiin varmuudella eri henkilöt.

Taulukko 10. *Keskeisimmät haasteet ja kehitysehdotukset toistuvuuksien mukaan.*

Haasteet ja kehitysehdotukset	**Toistuvuudet ulottuvuuksissa**
6x Markkinointilupatietojen kohdalla on ollut haasteita (esim. tietyt valinnat eivät ole välittyneet oikein).	1x ajantasaisuus, 1x maine, 1x oikea-aikaisuus, 1x täydellisyys, 1x uskottavuus ja 1x virheettömyys

6x Sähköposteissa on ollut ongelmia (esim. toimimattomia ja virheellisiä sähköposteja).	1x ajantasaisuus, 1x esityksen johdonmukaisuus, 1x maine, 1x objektiivisuus, 1x uskottavuus ja 1x virheettömyys
6x Master datan avainarvoja (esim. MDM-ID) ei välttämättä löydy muista datoista.	4x helppokäyttöisyys, 1x asiaankuuluvuus ja 1x sopiva määrä
6x Master asiakasdataan on valittu asiakkaat tiettyjen kriteerien perusteella. Tarve olisi kaikkien asiakkaiden datoille.	3x sopiva määrä, 1x asiaankuuluvuus, 1x objektiivisuus, 1x täydellisyys
6x Kehitettävää olisi tiettyjen kanavien alkuvaiheen rekisteröintiprosesseissa. Etenkin kevyen tunnistautumisen lähteistä on havaittu tulevan huonolaatuista dataa.	2x virheettömyys, 2x täydellisyys, 1x asiaankuuluvuus ja 1x maine
5x Tuotetietojen kohdalla on ollut haasteita (esim. ne eivät ole pitäneet paikkaansa).	2x asiaankuuluvuus, 1x maine, 1x uskottavuus ja 1x virheettömyys
4x Liiketoimintalogiikka ei ole täysin aukoton, kun yhdistetään eri asiakasnumeroiden tietoja yhdeksi riviksi.	2x maine, 1x esityksen johdonmukaisuus ja 1x virheettömyys
4x Datan prosessoinnin ja attribuuttien merkityksien dokumentointi on puutteellista.	4x tulkittavuus/ymmärrettävyys
3x Datalähteessä on tyhjiä, tyhjään viittaavia ja virheellisiä arvoja, minkä vuoksi niitä on myös Master datassa.	2x täydellisyys ja 1x virheettömyys
3x Yrityksen prosesseista johtuvia ongelmia, jos Master datasta puuttuu arvoja tai ne ovat virheellisiä.	2x virheettömyys ja 1x täydellisyys
2x Katuosoite voitaisiin pilkkoa yksityiskohtaisempiin attribuutteihin (esim. kadunnimi, kadunnumero, rappu, asunnon numero).	1x esityksen ytimekkyys ja 1x täydellisyys
2x Yrityksen tasolla ei ole selviä käytäntöjä attribuuttien nimeämisissä.	1x helppokäyttöisyys ja 1x tulkittavuus/ymmärrettävyys
2x Puuttuu keinoja ajantasaisuuden varmistamiseen.	2x ajantasaisuus

2x Datan lataukset voisi tapahtua useammin kuin kerran päivässä.	2x ajantasaisuus
2x Master datassa on testiasiakkaita, jotka olisi siivottava sieltä pois.	1x uskottavuus ja 1x virheettömyys

Haastatteluissa nousi esiin kuusi kertaa viisi eri haastetta tai kehitysehdotusta. Markkinointilupiin liittyviä haasteita ilmeni, vaikka haastatteluissa keskityttiin Master asiakasdatan yhteystietoihin. Toisaalta markkinointiluvat ovat oleellisesti sidoksissa asiakasdataan ja asiakkaiden toimintaan. Sähköpostien ongelmat ilmenivät haastatteluissa myös kuusi kertaa, joista suositteluautomaation työntekijöiden vastauksia oli viisi. Suositteluautomaation työntekijöiden tehtävissä oikeat ja toimivat sähköpostit ovat keskeisessä osassa, koska he lähettävät markkinointiviestejä. Master datassa todettiin olevan hyviä avainarvoja, joita ei välttämättä löydy kuitenkaan muista datoista. Toiveena oli, että avainarvoja hyödynnettäisiin enemmän eri liiketoimintojen datoissa, jotta esimerkiksi datan yhdistettävyys parantuisi. Master asiakasdatan sopivaan määrään vaikutti etenkin se, kun siihen on valittu asiakkaat tiettyjen kriteerien perusteella. Toisaalta Master datan voidaan todeta olevan vielä jonkinasteisessa kehitysvaiheessa, minkä vuoksi siihen on valittu tietyt asiakastyypit. Kehityskohteena tunnistettiin myös alkuvaiheen rekisteröintiprosessit, sillä eri kanavista voi tulla eritasoista dataa. Esimerkiksi joissain kanavissa vaaditaan asiakkailta kaikki tiedot, kun taas tietyissä kanavissa saattaa riittää ainoastaan sähköposti.

Tuotetietojen haasteet esiintyivät haastatteluissa viisi kertaa. Haastateltavat eivät täysin luottaneet Master datan tuotetietoihin, minkä vuoksi niitä ei juurikaan hyödynnetty. Master datan liiketoimintalogiikkaan liittyvät haasteet nousivat haastatteluissa esiin neljä kertaa. Eri asiakasnumeroiden tietoja yhdistetään yhdeksi riviksi, minkä vuoksi voi olla vaikea päätellä, mitä arvoa asiakas toivoo käytettävän. Lisäksi Master asiakasdatan toivottiin olevan joustavampaa, jotta asiakasattribuuteilla voisi olla useampia arvoja. Haastatteluissa ilmeni neljä kertaa myös dokumentoinnin puutteellisuus. Ei esimerkiksi tunnettu, miten Master dataa kerätään ja miten sitä prosessoidaan. Lisäksi kaikki attribuutit eivät olleet helposti ymmärrettävissä, minkä vuoksi kaivattiin metadataa attribuuttien merkityksien selventämiseksi.

Haastatteluissa esiintyi kolme kertaa, että datalähteessä on tyhjiä ja virheellisiä arvoja, minkä vuoksi niitä on myös Master datassa. Tähän liittyy vahvasti myös toinen kolme kertaa esiintynyt asia, mikä viittaa yrityksen prosessiongelmiin, jos Master datasta puuttuu arvoja tai ne ovat virheellisiä. Molemmat linkittyvät jo aiemmin esitettyyn alkuvaiheen rekisteröintiprosessien kehittämiseen, jotta asiakkaista saataisiin mahdollisimman täydelliset ja oikeat tiedot. Toisaalta muutkin prosessiongelmat voivat aiheuttaa datapoikkeamia.

Haastatteluissa esiintyi kaksi kertaa toive pilkotummalle katuosoitteelle. Se voitaisiin jakaa yksityiskohtaisempiin attribuutteihin, kuten kadunnimeen, kadunnumeroon, rappuun ja asunnon numeroon. Kaksi kertaa esiintyi myös attribuuttien nimeämisten epäselvyys, sillä esimerkiksi eri liiketoimintayksiköt saattavat nimetä samaa tarkoittavan attribuutin eri tavalla. Ajantasaisuuden kohdalla todettiin, että sen varmistamiseen puuttuu keinoja. Ei oltu varmoja, kuinka ajantasaista Master asiakasdata lopulta on. Suositteluautomaation työntekijöiden osalta ilmeni tarve useammin tapahtuville datan latauksille, jolloin voitaisiin reagoida nopeammin asiakkuuksissa tapahtuviin muutoksiin. Viimeinen kaksi kertaa esiintynyt asia oli Master asiakasdatassa olevat testiasiakkaat, jotka vaikuttavat datan uskottavuuteen ja virheettömyyteen.

6.3 Objektiivisen mittaamisen tulokset

Ajantasaisuutta mitattiin vertaamalla vanhoiksi luokiteltuja asiakasrivejä kokonaislukumäärään. Vanhoiksi riveiksi luokiteltiin kaikki ne, joiden viimeisimmästä päivityksestä oli yli 2 vuotta aikaa. Data kerättiin 11.01.2019, joten data oli vanhaa, jos viimeisin päivitys oli tapahtunut ennen 11.01.2017. Taulukossa 11 on esitetty ajantasaisuuden objektiivisen mittarin tulokset.

Taulukko 11. *Ajantasaisuuden objektiivisen mittaamisen tulokset.*

Ulottuvuus	Ajantasaisuus
Mittausmenetelmä	Verrataan vanhoiksi luokiteltuja rivejä kokonaislukumäärään
Mittari	1 - (Viimeisestä päivityksestä on yli 2 vuotta/Kokonaislukumäärä)
Tulos	0,597

Ajantasaisuuden objektiivisen mittarin tulokseksi saatiin 0,597. Datasta ei saatu selville, mitä attribuuttia ollaan muutettu päivityksen yhteydessä. Tämän vuoksi ajantasaisuuden mittari ei ole attribuuttikohtainen, vaan asiakasrivikohtainen. Toisaalta on hyvä huomioida, että asiakkaan tiedot voivat olla ajantasaisia, vaikka niitä ei olla päivitetty yli kahteen vuoteen. Ajantasaisuuden haasteellisuuteen liittyy myös asiakkaiden tietojen tarkistaminen asiakasrajapinnoissa, mistä ei välttämättä jää merkintää dataan.

Täydellisyyttä mitattiin vertaamalla tyhjiä ja tyhjään viittaavia arvoja kokonaislukumäärään. Tyhjään viittaavat arvot olivat attribuuttikohtaisia. Nimen kohdalla niitä olivat ”-” ja ”* *”, sähköpostin kohdalla ”0”, ”-”, ”--”, ”null@null”, ”eiole@”, ”eisahkopostia@” ja ”eipostia@”, katuosoitteen kohdalla ”0”, ”00” ja ”--” sekä postinumeron kohdalla ”0”, ”00”, ”000”, ”0000” ja ”00000”. Sähköpostien tyhjään viittaavien arvojen tunnistaminen

on haasteellista, sillä asiakkaalla voi olla toimiva ”eisahkopostia@gmail.com” -sähköposti. Yleisesti ottaen kyseiset tapaukset kuitenkin viittaavat tyhjään arvoon. Henkilötunnuksessa ja puhelinnumerossa ei ollut tyhjien arvojen lisäksi tyhjään viittaavia arvoja. Taulukossa 12 on esitetty objektiivisen mittaamisen tulokset.

Taulukko 12. *Täydellisyyden objektiivisen mittaamisen tulokset.*

Ulottuvuus	Täydellisyys
Mittausmenetelmä	Verrataan tyhjiä ja tyhjään viittaavia arvoja kokonaislukumäärään
Mittari	1 - (Tyhjät ja tyhjään viittaavat arvot/Kokonaislukumäärä)
Nimi	1
Henkilötunnus	0,908
Puhelinnumero	0,816
Sähköposti	0,879
Katusoite	0,998
Katuosoite, postinumero ja postitoimipaikka	0,956
Kaikki yhteystiedot	0,763

Nimissä oli vain 14 tyhjää ja tyhjään viittaavaa arvoa, minkä myötä sen pyöristetty tulos oli 1 eli parhain mahdollinen. Attribuuttien tasolla alhaisin tulos (0,816) oli puhelinnumeroilla. Yksittäisten attribuuttien täydellisyyden lisäksi mitattiin myös katuosoitteen, postinumeron ja postitoimipaikan sekä kaikkien yhteystietojen täydellisyyden tasoa. Data kerättiin hyödyntämällä postitoimipaikkojen suodatusta, minkä vuoksi siinä ei ollut tyhjiä ja tyhjään viittaavia arvoja. Postinumeroissa oli epätäydellisiä arvoja, minkä myötä kaikkien osoiteattribuuttien täydellisyys sai hieman alhaisemman tuloksen katuosoitteen täydellisyyteen verrattuna. Kaikkien yhteystietoattribuuttien täydellisyyden tasoksi saatiin 0,763.

Virheettömyyden mittaamisessa hyödynnettiin kohdeyrityksen liiketoimintasääntöjä, jotka tarkastelevat arvojen validiutta niiden muodon ja sisällön näkökulmista. Virheettömyyden mittarissa verrataan epävalideja arvoja kokonaislukumäärään. Taulukossa 13 on esitetty virheettömyyden objektiivisen mittaamisen tulokset.

Taulukko 13. *Virheettömyyden objektiivisen mittaamisen tulokset.*

Ulottuvuus	Virheettömyys
Mittausmenetelmä	Verrataan muodoltaan ja sisällöltään epävalideja arvoja kokonaislukumäärään
Mittari	1 - (Epävalidit arvot/kokonaislukumäärä)
Nimi	0,998
Henkilötunnus	0,905
Puhelinnumero	0,815
Sähköposti	0,876
Katusoite	0,962
Kaikki yhteystiedot	0,748

Kuten täydellisyydenkin kohdalla, nimi sai korkeimman tuloksen (0,998) ja puhelinnumero alhaisimman (0,815). Huomioitavaa on, että liiketoimintasäännöt ottavat huomioon myös tyhjät ja tyhjään viittaavat arvot. Tämän vuoksi epätäydelliset arvot ovat myös virheellisiä. Täydelliset arvot voivat olla virheellisiä, minkä vuoksi virheettömyyden tulokset ovat hieman alhaisempia kuin täydellisyyden.

Esitettyjen kolmen eri ulottuvuuden mittareiden lisäksi tarkasteltiin myös sähköpostien oikeellisuutta. Master asiakasdatasta nähtiin, jos asiakas oli vahvistanut sähköpostin. Vahvistettua sähköpostia voidaan pitää oikeana ja toimivana sähköpostina. Oikeellisuutta mitattiin vertaamalla vahvistamattomia sähköposteja kokonaislukumäärään. Taulukossa 14 on esitetty oikeellisuuden objektiivisen mittaamisen tulokset.

Taulukko 14. *Oikeellisuuden objektiivisen mittaamisen tulokset.*

Ulottuvuus	Oikeellisuus
Mittausmenetelmä	Verrataan asiakkaalta vahvistettuja sähköposteja kokonaislukumäärään
Mittari	1 – (Vahvistamattomat sähköpostit/Kokonaislukumäärä)
Sähköposti	0,312

Sähköpostien oikeellisuuden tulokseksi saatiin 0,312. Suurin osa tarkastelun kohteena olevasta datasta oli vahvistamattomia sähköposteja. Tähän liittyy myös haastatteluissa esiintynyt kehitysehdotus, jonka mukaan kaikki sähköpostiosoitteet tulisi vahvistaa asiakkaalta niiden oikeellisuuden ja toimivuuden varmistamiseksi.

6.4 Vertailuanalyysin tulokset

Vertailuanalyysiä varten subjektiivisen mittaamisen tulokset kerrottiin 0,1:llä, jotta ne saatiin välille 0-1. Objektiivisen mittaamisen tulokset puolestaan pyöristettiin kahden desimaalin tarkkuudelle, jolloin molempien mittausmenetelmien tulokset olivat samalla tarkkuudella. Täydellisyyden ja virheettömyyden objektiivisen mittaamisen tuloksissa huomioitiin kaikkien yhteystietoattribuuttien tulokset. Taulukossa 15 on esitetty subjektiivisen ja objektiivisen mittaamisen tulokset.

Taulukko 15. *Subjektiivisen ja objektiivisen mittaamisen tulokset.*

Ulottuvuus	Subjektiivinen mittaaminen	Objektiivinen mittaaminen
Ajantasaisuus	0,85	0,60
Täydellisyys	0,83	0,76
Virheettömyys	0,74	0,75

Ajantasaisuuden kohdalla subjektiivisen ja objektiivisen mittaamisen tuloksien eroavaisuus oli kaikista suurin (0,25). Siihen varmasti vaikuttaa objektiivisen mittarin määritys, jossa vanhaksi dataksi luokiteltiin ne rivit, joiden viimeisimmästä päivityksestä oli yli kaksi vuotta aikaa datan keräämisajankohtaan nähden. Data voi olla myös ajantasaista, vaikka viimeisimmästä päivityksestä olisi yli 2 vuotta aikaa. Haastatteluissa ilmeni, että Master asiakasdata vaikuttaa riittävän ajantasaiselta, mutta toisaalta puuttuu keinoja sen varmistamiseen. Lopulta voidaan todeta, että tietyn ajankohdan perusteella voi olla vaikea luokitella dataa vanhaksi.

Täydellisyydessä mittausmenetelmien tuloksien ero oli suhteellisen pieni (0,07). Haastatteluissa datalähteessä todettiin olevan tyhjiä ja tyhjään viittaavia arvoja, minkä vuoksi niitä on myös Master datassa. Toisaalta tyhjien ja tyhjään viittaavien arvojen lisäksi tuotiin esille tarvetta uusille attribuuteille ja datoille, yksityiskohtaisemmille osoiteattribuuteille sekä sieltä todettiin puuttuvan osa asiakkaista. Jos haastatteluissa tarkastelua oltaisiin rajattu ainoastaan tyhjään ja tyhjään viittaaviin arvoihin, niin tuloskin olisi voinut olla erilainen.

Virheettömyyden kohdalla subjektiivisen ja objektiivisen mittaamisen tuloksien eroavaisuus oli kaikista pienin (0,01). Virheettömyyden objektiiviset tulokset perustuvat kohdeyrityksen liiketoimintasääntöihin. Voi olla vaikea määritellä liiketoimintasääntöjä, jotka osaavat tunnistaa virheellistä ja epätäydellistä dataa absoluuttisesti. Haastatteluissa puolestaan nousi esiin etenkin alkuvaiheen prosessien rooli virheellisen datan tuottamisessa.

7. PÄÄTELMÄT

Tässä luvussa esitetään ensin yleisiä päätelmiä teorian ja empirian vertailun muodossa. Tämän jälkeen pureudutaan tarkemmin tulosten päätelmiin, tutkimuskysymysten vastauksiin ja kehitysehdotusten esittämiseen.

Datan laadun menetelmien vaiheita voidaan yleisesti jakaa kolmeen osaan; tilan rekonstruktioon, mittaukseen ja arviointiin sekä kehitystoimenpiteisiin (Batini et al. 2009). Tässä työssä keskityttiin näistä vaiheista pääosin arviointiin ja mittaukseen. Tilan rekonstruktiota ei koettu tarpeelliseksi, johtuen nykyisten laatuongelmien tiedostamisesta erilaisten työtehtävien myötä. Erilaisia kehitystoimenpiteitä tuotiin teoriassa pintapuolisesti esille ja empirian perusteella esitetään myös muutamia konkreettisia kehitystoimenpiteitä luvussa 7.3. Kehitystoimenpiteet jätetään kuitenkin pääosin kohdeyrityksen harkinnan varaan.

Datan laatukirjallisuudessa on esitetty useita erilaisia objektiivisia mittareita, mutta niiden käytännön hyödyntäminen saattaa olla haasteellista. Esimerkkinä ajantasaisuuden mittaaminen, missä voi olla vaikea tietää datan toimitusaika, syöttöaika ja ikä. Suositeltavampaa voisi olla ymmärtää erilaisten mittareiden perusperiaatteet ja kehittää niiden pohjalta yrityksen toimintaan käytännössä sopivat mittarit. Tällöin kyse on objektiivisesta tehtävästä riippuvaisesta mittaamisesta, joka sisältää esimerkiksi yrityksen liiketoimintasäännöt (Pipino et al. 2002). Keskeistä on myös tuntea tarkastelun kohteena oleva data, minkä myötä tunnistetaan mahdolliset laatuongelmat ja pystytään määrittämään yksityiskohtaiset mittaamistarpeet. Empiriassa mittareita määrittäessä tunnettiin kohdeyrityksen Master asiakasdata varsin hyvin, mikä auttoi esimerkiksi sisällyttämään virheettömyyden mittariin yrityksen liiketoimintasäännöt.

Mittaamisessa on huomioitava, että kaikkia ulottuvuuksia ei voida mitata objektiivisesti. Esimerkiksi maine, tulkittavuus, ymmärrettävyys, uskottavuus ja asiaankuuluvuus vaativat käyttäjäkyselyiden toteuttamista (Cappiello et al. 2004; Watts et al. 2009). Toisaalta käyttäjäkyselyiden sijaan voi olla myös mahdollista toteuttaa muita menetelmiä. Esimerkiksi Batini et al. (2009) ehdottaa, että tulkittavuutta voidaan tarkastella myös datan dokumentaation laadukkuuden näkökulmasta. Tämän työn empiriassa objektiiviseen mittaamiseen valittiin ulottuvuudet laatuongelmien, yrityksen tavoitteiden ja objektiivisen mittaamisen mahdollisuuksien mukaisesti. Mittarit perustuivat teoriassa esitettyyn yleiseen kaavaan, minkä myötä yksityiskohtaisemmat mittarit jäivät hyödyntämättä.

Yrityksillä voi olla kiinnostus yhden datan laatumittarin määrittämiseen, mistä näkisi nopeasti laadun yleistason. Tätä näkökulmaa puoltavia teoksia ei juurikaan lähdekirjallisuudesta löytynyt, vaan keskeistä on huomioida merkitykselliset ulottuvuudet ja kehittää mit-

tausmenetelmät kyseisille ulottuvuuksille (Wang et al. 1995; Wang & Strong 1996; Bronselaer et al. 2018b). Toisaalta yritykset voivat saada arvoa yksiarvoisesta koostemittarista, jos sitä tulkitaan oletusten ja rajoitusten mukaisesti (Pipino et al. 2002). Koostemittarista on vaikea havaita, missä ulottuvuuksissa, attribuuteissa tai datariveissä on laatuongelmia. Esimerkiksi empiriassa mahdollisuuksien mukaan toteutettu attribuuttikohtainen mittaaminen antaa yritykselle yksityiskohtaisempaa ymmärrystä laadun tasosta. Yhden ulottuvuuden mukainen mittari saattaa antaa myös vääristynyttä tietoa mittaamisen tavoitteeseen nähden. Esimerkiksi osoitetietojen hyödyllisyyttä tarkasteltaessa on tärkeää mitata useampia ulottuvuuksia. Osoitetiedot voivat olla täydellisiä, mutta postitoimipaikan ja postinumeron väliset epäjohdonmukaisuudet voivat tehdä tiedoista hyödyttömiä.

Työssä hyödynnetty Hybridi-arviointimenetelmä valittiin etenkin sen joustavuuden vuoksi, sillä se on mahdollista toteuttaa hyvinkin yksinkertaisesta arvioinnista yksityiskohtaiseen ja laajaan arviointiprojektiin. Empiriassa menetelmään valittiin kaikki ne arviointitoiminnallisuudet, jotka sopivat yrityksen tavoitteisiin ja tarpeisiin sekä samalla diplomityön laajuuteen. Hybridi-arviointimenetelmän toiminnallisuuksien läpikäyminen antaa myös kohdeyritykselle valmiudet sen laajempaan hyödyntämiseen. Kyseinen arviointimenetelmä tuntui toimivan melko hyvin kohdeyrityksen ja tämän työn kontekstissa, sillä sen hyödyntäminen oli suoraviivaista. Toisaalta Hybridi-arviointimenetelmä teoreettisesti perustui vain yhteen artikkeliin, mikä osaltaan voi vähentää sen tieteellistä luotettavuutta.

Arvioinnin yhtenä tavoitteena oli mitata tiettyjä tunnistettuja datan laatuongelmia objektiivisten ja subjektiivisten mittauksien avulla, minkä voidaan todeta toteutuneen. Toisena tavoitteena oli toteuttaa datan laadun arviointia ja mittaamista jatkuvan parantamisen periaatteiden mukaisesti. Arvioinnissa hyödynnettyä Hybridi-menetelmää voi käyttää eri datoihin tietyin aikavälein, mikä mahdollistaa jatkuvan parantamisen periaatteiden noudattamisen. Empiriassa käytettyjä mittareita ei voi puolestaan suoraan hyödyntää muihin datoihin, johtuen epämieluisten tulosten sisällön muuttumisesta. Samaan Master asiakasdataan kyseisiä mittareita voi käyttää suuntauksien seuraamiseen, jos tunnistetaan epämieluisat tulokset tarkasteltavien attribuuttien kohdalla.

7.1 Tulosten päätelmät

SUBJEKTIIVINEN MITTAAMINEN

Subjektiivisesta mittaamisesta saatiin määrällisen numeerisen aineiston lisäksi myös laadullista aineistoa. Numeeriset vastaukset olivat yllättävän korkealla tasolla, sillä asteikolla 0-10 alhaisin tulos työntekijäryhmien keskiarvon mukaan oli 6,50 ja kokonaiskeskiarvon mukaan 7,4. Tässä korostuu haastatteluissakin esiintynyt asia, jonka mukaan Master data edustaa kohdeyrityksen parasta dataa. Huolestuttavaa olisi, jos Master datan laatu olisi hyvin alhaisella tasolla.

Numeerisista vastauksista nähtiin, että virheettömyys, maine, objektiivisuus ja täydellisyys saivat alhaisimmat tulokset. Maineen tuloksiin todennäköisesti vaikutti MDM-projektiin ladatut korkeat odotukset, jotka eivät ainakaan kaikkien työntekijöiden kohdalla olleet toteutuneet. Toisaalta Master data ei vielä ollut lopullisessa, viimeistellyssä muodossa, joten maineen voisi olettaa kehittyvän MDM-projektin etenemisen myötä. Objektiivisuuden tuloksiin puolestaan vaikutti Master datan dokumentoinnin puutteellisuus, koska ei tiedetty datan keräysprosessia. Virheettömyys ja täydellisyys linkittyvät vahvasti asiakkaiden alkuvaiheen rekisteröintiprosesseihin. Jos rekisteröintikanavassa ei vaadita asiakkaalta kaikkia tietoja tai tietoja ei validoida syöttöhetkellä, niin ne näkyvät virheellisinä tai tyhjinä arvoina Master datassa.

Subjektiivisen mittaamisen tuloksissa korostui osittain työntekijäroolien vaikutus. Esimerkiksi datatieteilijöille Master asiakasdata oli riittävän ajantasaista ja oikea-aikaista, koska heille kyseinen data riittää taustatiedoiksi. Suositteluautomaation työntekijöiden kohdalla puolestaan nousi esiin myös tarve useammin tapahtuville datalatauksille, koska ajantasaiset ja oikea-aikaiset tiedot ovat avainasemassa markkinointiviestintää tehtäessä.

Toistuvuuden mukaisesti esitetyt haasteet ja kehitysehdotukset (taulukko 10) korostavat keskeisimpiä ongelmakohtia. Näihin palataan tarkemmin luvussa 7.3, jossa esitetään kehitysehdotuksia tuloksien perusteella.

OBJEKTIIVINEN MITTAAMINEN JA VERTAILUANALYYSI

Objektiivisen mittaamisen tulokset olivat hieman alhaisemmat kuin subjektiivisen mittaamisen. Ainoastaan virheettömyyden tulos oli objektiivisen tarkastelun kohdalla hieman subjektiivista mittaamista korkeampi (0,01). Objektiivisen mittaamisen tuloksiin vaikuttavia tekijöitä ovat etenkin datan pieni otoskoko Master asiakasdatamassasta ja mittareiden määritys. Kokonaisuutena vertailuanalyysin tulokset olivat yllättävänkin lähellä toisiaan, ottaen huomioon esimerkiksi haastatteluiden vastauksien laajuudet tarkoitettuun rajaukseen nähden.

Haastatteluissa nousi esiin, että historiasta johtuen datasta voi löytyä erikoisia poikkeamia, datalähteissä on epätäydellisiä ja virheellisiä arvoja, sähköposteissa on ollut ongelmia sekä rekisteröintiprosesseissa olisi kehitettävää. Nämä haasteet näkyivät etenkin täydellisyyden, virheettömyyden ja oikeellisuuden objektiivisen mittaamisen tuloksissa. Sähköposteja ei vahvisteta asiakkailta kaikissa rekisteröintikanavissa, minkä vuoksi sen tulos oli selkeästi alhaisin (0,312). Mittareissa hyödynnettyjen liiketoimintasääntöjen mukaisesti epätäydelliset arvot ovat myös virheellisiä, minkä seurauksena virheettömyyden tulokset olivat hieman täydellisyyden tuloksia heikommat. Samalla täydellisyyden ja virheettömyyden tuloksien erosta voi suoraan nähdä, että datassa on epätäydellisten arvojen lisäksi myös virheellisiä arvoja.

Ajantasaisuuden mittaaminen oli selkeästi haastavinta, sillä tietyn ajankohdan mukaisesti mitattua ajantasaisuutta ei voida pitää täysin luotettavana. Keskeisin haaste tässä näkökulmassa on, että asiakkaiden tiedot voivat pysyä ajantasaisina hyvinkin pitkän aikaa. Lisähaasteita tuo, kun asiakasrajapinnoissa tehdyistä asiakkaiden tietojen tarkastuksista ei välttämättä jää merkintää dataan. Ajantasaisuuden osalta suuri vastuu on itse asiakkailla, joiden on päivitettävä tuoreimmat tietonsa yrityksen järjestelmiin. Yritys voi myös aktivoida asiakkaita omien tietojen päivittämiseen esimerkiksi erilaisilla kampanjoilla. Jos taas on saatavilla yksityiskohtaista tietoa datan toimitus- ja syöttöajasta, iästä sekä volatiliteetista, niin on mahdollista muodostaa luotettavampia mittareita ajantasaisuuden ja oikea-aikaisuuden mittaamiseen.

7.2 Tutkimuskysymysten vastaukset

Tutkimuksen tarkoituksena oli selvittää, miten kohdeyrityksen datan laatua voidaan mitata ja arvioida. Ensimmäinen alatutkimuskysymys oli:

- *Mitä datan laadulla tarkoitetaan?*

Tähän tutkimuskysymykseen vastattiin luvussa 2, tarkemmin alaluvussa 2.2. Määritelmä "sopivuus käyttötarkoitukseen" (*engl. Fitness for use*) on laajasti hyväksytty laatukirjallisuudessa. Datan laadulla tarkoitetaan siis sitä, miten hyvin se sopii datan kuluttajien käyttötarpeisiin. Keskeistä on myös huomioida, miten hyvin data esittää kuluttajien mielestä sitä, mitä se on tarkoituskin esittää. Datan laatua voidaan tarkastella esimerkiksi suunnittelun laadun ja vaatimustenmukaisuuden laadun näkökulmista. Datan laadun ulottuvuudet puolestaan esittävät yhtä datan laadun näkökulmaa tai rakennetta, mitkä mahdollistavat datan laadun mittaamisen ja hallinnan.

Tutkimuksen toinen alatutkimuskysymys oli:

- *Minkälaisia menetelmiä datan laadun mittaamiseen ja arviointiin on olemassa?*

Tähän tutkimuskysymykseen vastattiin luvuissa 3 ja 4. Yleisesti ottaen datan laadun mittaamisessa on otettava huomioon erilaiset ulottuvuudet ja kehitettävä mittausmenetelmät valituille ulottuvuuksille. Mittaamisella tarkoitetaan toimintoa, jossa määritetään numeroarvo tarkastelun kohteena olevalle attribuutille.

Datan laadun mittaaminen voi olla objektiivista (rakenteellista) tai subjektiivista (sisältöpohjaista). Objektiivinen mittaaminen perustuu datan fyysisiin ominaisuuksiin, kuten lukumäärien suhteisiin tai aikamittauksiin. Objektiiviset mittarit voivat olla tehtävästä riippumattomia (kuvaavat datan tilaa ilman asiayhteystietoa) tai tehtävästä riippuvaisia (sisältävät esim. yrityksen liiketoimintasäännöt). Subjektiivinen mittaaminen puolestaan perustuu datan käyttäjien mielipiteisiin ja se heijastaa käyttäjien tarpeita ja odotuksia. Taloudellisesta näkökulmasta objektiivinen mittaaminen voidaan liittää kustannuksiin ja

subjektiivinen mittaaminen datan laadun parantamisesta saataviin hyötyihin. Mittaaminen voi olla myös staattista tai dynaamista. Staattisessa mittauksessa mitataan tutkittavan datan tilannekuvaa, kun taas dynaamisessa mittaamisessa mitataan dataa sen virran tiettyjen kohtien aikana. Laadun seurantaa varten tulee valita muutamia keskeisimpiä mittareita, koska kaikkia mittareita ei voi eikä pitäisi seurata. Mittareiden valinnassa voidaan huomioida useita eri tekijöitä, kuten mittarin prioriteetti, mittausmenetelmä, mittaustiheys, kustannusten ja hyötyjen suhde sekä huomiotta jättämisen riski.

Datan profilointia voidaan myös hyödyntää datan laadun mittaamisessa. Se on tietyntyyppinen data-analyysi, jota käytetään datajoukon ominaisuuksien etsimiseen ja luonnehtimiseen. Tuloksena saadaan tietoa datan ominaisuuksista, kuten datan tyypeistä, kentän pituuksista, arvojoukoista, formaatti- ja sisältömalleista sekä epäsuorista säännöistä. Datan profiloinnin keskeisimmät menetelmät voidaan jakaa kolmeen ryhmään, jotka ovat rakenteen, sisällön ja suhteiden analysointi.

Datan laadun arvioinnilla tarkoitetaan prosessia, jossa hyödynnetään datan laadun mittauksia laadun diagnosoimiseksi ja tarvittavien datan laadun kehittämistoimenpiteiden määrittämiseksi. Mittauksen keskeisimpänä tarkoituksena on määrällisen merkityksen tarjoaminen siitä, kuinka monessa laadun ulottuvuudessa päästään tavoitteeseen. Yleisimmissä tapauksissa arviointimenetelmät koostuvat kolmesta vaiheesta, jotka ovat tilan rekonstruktio, mittaus sekä kehittämistoimenpiteet. Tässä työssä esitettiin neljä erilaista arviointimenetelmää, jotka olivat TDQM-, AIMQ-, DQA-, ja Hybridi-menetelmä. Datan laatukirjallisuudessa on esitetty myös useita muita arviointimenetelmiä, mutta niiden perusperiaatteet voivat olla hyvinkin samankaltaisia.

Viimeinen ja samalla kokoava alatutkimuskysymys oli:

- *Miten data, laatu, mittaaminen ja arviointi liittyvät toisiinsa?*

Tiivistetysti voidaan todeta, että kyseiset termit liittyvät toisiinsa kokonaisvaltaisen datan laadun diagnosointi- ja kehitysprojektin toteuttamiseksi. Jokaisen termin sisältö on ymmärrettävä, jotta datan laadun arviointiprojektin toteuttaminen on mahdollista. Projektin aluksi on määriteltävä, mitä arviointimenetelmää hyödynnetään ja minkälaisia arviointitoiminnallisuuksia siihen sisällytetään. Tämän lisäksi määritetään tarkastelun kohteena oleva data. Tarkasteltavana datana voi olla esimerkiksi Master data tai transaktiodata, ja vielä tarkemmin asiakas-, tuote- tai sopimusdata. Lisäksi voidaan vielä kohdentaa, minkä tietojärjestelmän data tai mitkä attribuutit projektiin valitaan.

Datan laadun kohdalla on tärkeää tiedostaa, mikä laadun näkökulma otetaan huomioon. Samalla on selvitettävä, minkälaisia laatuongelmia on havaittu ja mikä on tavoiteltu laadun taso. Laatuongelmien ja tavoitteiden selvittämisen myötä pystytään valitsemaan niitä vastaavat laadun ulottuvuudet, jotka toimivat perustana datan laadun mittaamiselle.

Datan laadun mittaamisessa keskeistä on valita mittausmenetelmä, joka pääsääntöisesti voi olla objektiivista tai subjektiivista. Objektiivisessa mittaamisessa valitaan tai luodaan ulottuvuuksille mittarit, joiden valinnassa voidaan huomioida useita eri tekijöitä. Mittareiden hyödyntäminen ei ole kertaluontoista, vaan tarkoituksena on seurata laadun kehittymistä ja mahdollisten poikkeavuuksien ilmentymistä. Subjektiivisessa mittaamisessa määritetään menetelmä, jolla kerätään mielipiteitä laadun tasosta kyseiseen dataan olennaisesti liittyviltä henkilöiltä. Mahdollista on myös verrata eri mittausmenetelmien tuloksia keskenään kokonaisymmärryksen muodostamiseksi.

Datan laadun arvioinnissa hyödynnetään mittauksien arvoja laadun diagnosoimiseksi ja samalla myös datan laadun kehittämistoimenpiteiden määrittämiseksi. Kehittämistoimenpiteissä voidaan hyödyntää kahta erilaista strategiaa, data- ja prosessipohjaista strategiaa. Datapohjaiset strategiat parantavat datan laatua suoraan muuttamalla datan arvoja, kun taas prosessipohjaiset strategiat parantavat datan laatua uudelleensuunnittelemalla prosesseja, jotka luovat tai muokkaavat dataa.

Lopulta diplomityön päätutkimuskysymys oli:

- *Miten kohdeyrityksen datan laatua voidaan mitata ja arvioida?*

Päätutkimuskysymykseen saatiin vastaus alatutkimuskysymyksien vastauksien avulla. Voidaan kuitenkin vielä todeta, että kohdeyrityksen datan laatua voidaan mitata ja arvioida toteuttamalla arviointiprojekti, johon sisällytetään laatuongelmia ja -tavoitteita vastaavat toiminnot. Tutkimuksen toteuttamisen myötä nähtiin, että Hybridi-arviointimenetelmä on toimiva vaihtoehto datan kokonaisvaltaiseksi arviointiprojektiksi.

7.3 Kehitysehdotukset

Taulukossa 16 on esitetty kehitysehdotuksia haastatteluiden perusteella. Kehitysehdotuksien lisäksi on esitetty, mihin laadun ulottuvuuksiin niillä on mahdollista vaikuttaa.

Taulukko 16. *Kehitysehdotukset ja niiden vaikutukset laadun ulottuvuuksiin.*

Kehitysehdotus	Laadun ulottuvuudet
MDM-ID:n hyödyntäminen yrityksen tasoisesti	Asiaankuuluvuus, esityksen johdonmukaisuus, helppokäyttöisyys, maine, saatavuus, tulkittavuus/ymmärrettävyys
Yhdenmukaiset attribuuttien nimeämiskäytännöt	Esityksen johdonmukaisuus, helppokäyttöisyys, maine, tulkittavuus/ymmärrettävyys

Datan dokumentaation parantaminen	Helppokäyttöisyys, maine, tulkittavuus/ymmärrettävyys, uskottavuus
Alkuvaiheen rekisteröintiprosessien kehittäminen	Asiaankuuluvuus, maine, täydellisyys, virheettömyys
Sähköpostien vahvistaminen kaikissa kanavissa	Ajantasaisuus, asiaankuuluvuus, maine, oikeellisuus, virheettömyys
Katuosoite yksityiskohtaisempiin attribuutteihin	Asiaankuuluvuus, esityksen ytimekkyys, helppokäyttöisyys, maine, tulkittavuus/ymmärrettävyys
Master datan liiketoimintalogiikan kehittäminen	Asiaankuuluvuus, helppokäyttöisyys, maine

MDM-ID:n kokonaisvaltainen hyödyntäminen helpottaa etenkin datojen yhdistettävyyttä ja ymmärrettävyyttä sekä vähentää epäselvyyksiä. Laajemman hyödyntämisen myötä levitetään tietoisuutta Master datan hyödyistä, minkä lisäksi MDM-ID:tä on myös tietoturvallisempi käyttää kuin esimerkiksi henkilötunnusta.

Haastatteluissa nousi esiin, että esimerkiksi Master datasta luoduissa näkymissä on saatettu nimetä attribuutteja eri tavalla lähdejärjestelmään nähden. Attribuuttien nimeämiskäytäntöjen tulisi olla selkeämmät, sillä yhdenmukaisilla nimeämiskäytännöillä on suurin vaikutus datojen tulkittavuuteen ja ymmärrettävyyteen. Attribuuttien nimeämiskäytäntöjen yhdenmukaistamiseen liittyy vahvasti myös dokumentaation kehittäminen. Monet haastateltavat eivät tienneet, miten kyseistä dataa prosessoidaan ja millä logiikalla datan arvot valitaan asiakkaille. Datan dokumentaatiossa olisi hyvä olla esimerkiksi data standardit, datan virtauskaavioita ja datan mallinnuksia.

Alkuvaiheen rekisteröintiprosessit tunnistettiin melko isoksi kehityskohteeksi. Haastatteluissa nousi esiin, että joissain kanavissa vaaditaan asiakkailta kaikki tiedot, kun taas tietyissä kanavissa saattaa riittää ainoastaan sähköposti. Olisikin löydettävä keino oleellisten asiakastietojen keräämiseen ilman myyntiputken hidastamista. Samojen tietojen kerääminen kaikissa kanavissa parantaa lähdejärjestelmän laatua (etenkin täydellisyyttä), mikä näkyy lopulta Master datan laadun kehittymisenä. Suurin vaikutus rekisteröintiprosessien laadun kehittämiseen on tietojen syöttöhetken validoinnilla, mikä estää virheellisen datan virtaamisen tietojärjestelmiin.

Asiakkailta ei vahvistettu sähköposteja kaikissa kanavissa, mikä näkyi sähköpostien oikeellisuuden heikkona tuloksena. Sähköpostien vahvistaminen asiakkailta on keskeinen tapa varmistua sähköpostien oikeellisuudesta ja toimivuudesta. Sähköpostien vahvistaminen voidaan toteuttaa esimerkiksi lähettämällä vahvistuslinkki asiakkaan sähköpostiin.

Oikeat ja toimivat sähköpostit tehostaisivat markkinointiviestintää toteuttavia työntekijäryhmiä, kuten suositteluautomaation työntekijöitä.

Haastatteluissa nousi esiin tarve pilkotummalle katuosoitteelle. Yksityiskohtaisemmat osoiteattribuutit helpottaisivat datan käsittelyä ja niiden avulla olisi myös selkeämpää tarkastella osoitetietojen täydellisyyttä ja virheettömyyttä.

Haastatteluissa todettiin, että Master datan liiketoimintalogiikka ei ole aukoton, kun yhdistetään asiakkaiden tietoja yhdeksi riviksi. Datalta toivottiin tietynasteista joustavuutta, jotta samasta attribuutista voisi olla myös useampia arvoja. Keskeistä olisi mahdollistaa Master asiakasdataan esimerkiksi yhden puhelinnumeron lisäksi myös työnumeron sisällyttäminen.

8. POHDINTA

Tässä luvussa esitetään tutkimuksen yleisen pohdinnan lisäksi myös tutkimuksen arviointia, rajoituksia ja mahdollisia jatkotutkimusehdotuksia. Tutkimuksen yhtenä keskeisimpänä vaikutuksena voidaan pitää ymmärryksen kasvattamista datan laadun mittaamisen ja arvioinnin perusperiaatteista sekä menetelmistä. Suomenkielisten tutkimuksien vähäisyys kyseisestä aiheesta osaltaan myös nostattaa tutkimuksen arvoa.

Tutkimuksen aiheen keskeisen terminologian epäselvyys toi omat haasteensa. Osa lähdekirjallisuudesta erottaa mittaamisen (*engl. Measurement)* ja arvioinnin (*engl. Assessment*) toisistaan, kun taas osa saattaa puhua vain toisesta. Tämän tutkimuksen yhtenä tarkoituksena olikin erottaa kyseiset termit toisistaan, johtuen niiden erilaisista ominaispiirteistä.

Tutkimuksen tuloksista saatiin numeerisen laadun tason lisäksi myös laadullisia kehitystoimenpiteitä ja haasteita. Numeeriset tulokset antoivat yleiskuvan laadun tasosta eri ulottuvuuksien avulla, kun taas laadullisten tulosten myötä pystyttiin tunnistamaan konkreettisia ongelmakohtia. Tutkimuksen toteuttamisen myötä nähtiin myös, että Hybridi-arviointimenetelmä on potentiaalinen menetelmä datan laadun arviointiin etenkin siihen sisällytettävien arviointitoimintojen joustavuuden myötä.

8.1 Tutkimuksen arviointi

Tutkimuksen luotettavuutta voidaan arvioida validiteetin ja reliabiliteetin näkökulmista (Hirsjärvi et al. 2007 s. 226; Tuomi & Sarajärvi 2009 ss. 136-141). Validiteetti tarkoittaa, että tutkimuksessa on tutkittu luvattua asiaa, kun taas reliabiliteetti tarkoittaa tutkimustulosten toistettavuutta (Tuomi & Sarajärvi 2009 ss. 136-141). Reliabiliteetilla voidaan tarkoittaa, että tutkimus antaa ei-sattumanvaraisia tuloksia (Hirsjärvi et al. 2007 s. 226).

Tutkimuksen validiteetin näkökulmasta haastattelukysymykset pyrittiin muodostamaan helposti ymmärrettäviksi ja niitä myös selitettiin laajemmin haastatteluiden aikana. Haastattelukysymykset ovat myös tutkimuksen liitteenä ja objektiiviset mittarit on esitetty luvuissa 5.5 ja 6.3. Tutkimuksen teorian kirjallisuus perustuu datan laadun mittaamisen ja arvioinnin teoriaan, joten esitettyjen arviointi- ja mittausmenetelmien voidaan olettaa liittyvän tutkittavaan asiaan.

Tutkimuksessa hyödynnetty määrällinen Master asiakasdata edusti suhteellisen pientä otosta koko Master asiakasdatamassasta, mikä voi vaikuttaa tutkimuksen tuloksiin. Toisaalta otoksen valinnassa huomioitiin, ettei mikään tietty tunnettu datan laatuongelma korostu valitussa otoksessa. Haastateltavia olisi voinut olla enemmän ja useammista työntekijäryhmistä, mikä olisi kasvattanut tulosten luotettavuutta. Haastateltavien valintaan vaikutti Master asiakasdatan käyttäjien melko alhainen määrä, sillä Master asiakasdatan

voidaan todeta olevan vielä kehitysvaiheessa. Tulokset ovat myös vahvasti sidoksissa tutkimuksen toteuttamisajankohtaan, etenkin kasvavan datamäärän ja erilaisten laadun kehittämistoimenpiteiden vaikutusten vuoksi.

Validiteetti ja reliabiliteetti liittyvät kuitenkin pääosin määrälliseen tutkimukseen, eikä laadullisen tutkimuksen luotettavuuden arviointiin ole yksiselitteisiä ohjeita. Laadullisen tutkimuksen luotettavuutta voidaan arvioida esimerkiksi tutkimuksen kohteen ja tarkoituksen, tutkijan sitoutumisen, aineiston keruun, tutkimuksen tiedonantajien, tutkija-tiedonantaja-suhteen, tutkimuksen keston, aineiston analyysin sekä tutkimuksen raportoinnin näkökulmista. (Tuomi & Sarajärvi 2009 ss. 136-141) Laadullisen tutkimuksen luotettavuutta voidaan parantaa tarkalla selostuksella tutkimuksen toteuttamisesta (Hirsjärvi et al. 2007 s. 227). Tutkimuksen toteuttamisesta on kerrottu luvuissa 1.3 ja 5, mikä osaltaan lisää tutkimuksen luotettavuutta ja mahdollisuutta toistettavuuteen.

Tutkimuksessa hyödynnettiin triangulaatiota, jota voidaan käyttää liittyen tutkimuksen totuuden ongelmaan tai tutkittavan ilmiön kokonaisuuden hahmottamiseen. Kokonaisuuden hahmottamisella tarkoitetaan, että triangulaation tarkoituksena on saada tutkimukseen leveyttä ja syvyyttä. Triangulaatiota ei välttämättä voida käyttää tutkimuksen validiteettimenetelmänä, vaan enemmänkin monimuotoisten tulosten mahdollistajana ja tutkimuksen kiinnostavuuden kohottajana. (Tuomi & Sarajärvi 2009 ss. 144-149) Toisaalta Hirsjärvi et al. (2007 s. 228) kertovat, että tutkimuksen validiutta voidaan parantaa hyödyntämällä useita menetelmiä, triangulaatiota. Tässä tutkimuksessa triangulaatiota hyödynnettiin tutkittavan ilmiön kokonaisuuden hahmottamiseen, koska datan laadun tason kokonaisvaltaiseen määrittämiseen vaaditaan objektiivisia määrällisiä menetelmiä ja subjektiivisia laadullisia menetelmiä. Tämän myötä laadullisten ja määrällisten tiedonkeruu- ja analysointimenetelmien hyödyntäminen oli loogista.

8.2 Tutkimuksen rajoitukset

Tutkimuksen rajoituksena voidaan pitää etenkin objektiivisessa mittaamisessa hyödynnetyn Master asiakasdatan suhteellisen pientä otoskokoa. Suuremmassa määrällisen datan otoskoossa olisi todennäköisesti esiintynyt myös muita virheellisiä ja epätäydellisiä arvoja, jotka olisivat saattaneet vaikuttaa kokonaistuloksiin. Lisäksi empiriassa keskityttiin Master dataan, joka yleisesti ottaen edustaa yrityksen laadukkainta dataa. Suurempia laatupoikkeamia olisi voinut esiintyä esimerkiksi transaktiodataa tutkittaessa.

Tutkimuksen rajoituksena on myös keskittyminen datan arvoihin. Toisaalta empirian haastatteluissa nousi esiin myös esimerkiksi dataprosesseihin liittyviä asioita. Tämän lisäksi tutkimuksen painopiste oli datan laadun mittaamisessa ja arvioinnissa, minkä myötä kehitystoimenpiteet käsiteltiin vain pintapuolisesti. Tutkimusta oli kuitenkin rajattava, sillä triangulaation hyödyntäminen toi tutkimukseen huomattavasti laajuutta.

Tutkimuksen toteuttamishetkellä Master data oli vielä kehitysvaiheessa, mikä osaltaan saattoi vaikuttaa tuloksiin. Tietynasteisen kehitysvaiheen tutkimista voidaan kuitenkin pitää arvokkaana, sillä se antaa tietoa mahdollisista ongelmakohdista ennen siirtymistä kokonaisvaltaiseen Master datan hyödyntämiseen.

Subjektiivisessa mittaamisessa haastatteluita olisi voinut rajata tarkemmin. Haastatteluiden tarkoituksena oli keskittyä Master asiakasdataan, mutta usein vastaukset kohdistuivat koko Master dataan, sisältäen esimerkiksi tuotedatan. Laajemmat vastaukset kuitenkin huomioitiin, sillä niillä koettiin olevan arvoa kohdeyritykselle. Objektiivisessa mittaamisessa puolestaan tarkasteltiin vain Master asiakasdataa, minkä myötä näiden eri mittaamismenetelmien tuloksien vertailukelpoisuus heikentyi.

8.3 Jatkotutkimusehdotukset

Yhtenä jatkotutkimusehdotuksena on erilaisten datan laadun mittaamisen ja arvioinnin työkalujen käsitteleminen. Esimerkiksi datan profiloinnin tai objektiivisia mittareita sisältävien työkalujen toiminnallisuuksien läpikäyminen toisi yrityksille konkreettisia vaihtoehtoja laadun diagnosointiin. Mielenkiintoisena näkökulmana olisi myös tarkempi perehtyminen tiettyjen datatyyppien mittaus- ja arviointimenetelmiin. Esimerkiksi voisi selvittää, sopivatko jotkin mittausmenetelmät paremmin transaktiodatan kuin Master datan mittaamiseen.

Datan laadun kehittämistoimenpiteet käsiteltiin melko pintapuolisesti, minkä myötä niitä olisi mahdollisuus tutkia laajemmin. Etenkin empiriassa olisi tilaa käsitellä erilaisten data- ja prosessipohjaisten strategioiden hyödyntämistä datan laadun kehittämispyrkimyksissä. Muitakin datan laadun arviointimenetelmiä olisi tutkittavana, mutta toisaalta niiden perusperiaatteet noudattavat pitkälti samaa kaavaa. Tämän vuoksi tässä työssä käsiteltiin tarkemmin Hybridi-arviointimenetelmää, johon sisältyy myös muiden arviointimenetelmien toiminnallisuuksia.

LÄHTEET

Aljumaili, M., Karim, R. & Tretten, P. (2016). Metadata-based data quality assessment, VINE Journal of Information and Knowledge Management Systems, Vol. 46(2), pp. 232.

Andreescu, A.I., Belciu, A., Florea, A. & Diaconita, V. (2014). Measuring Data Quality in Analytical Projects, Database Systems Journal, (1), pp. 15-25.

Azeroual, O., Saake, G. & Schallehn, E. (2018). Analyzing data quality issues in research information systems via data profiling, International Journal of Information Management, Vol. 41 pp. 50-56.

Ballou, D.P. & Pazer, H.L. (2003). Modeling completeness versus consistency tradeoffs in information decision contexts, IEEE Transactions on Knowledge and Data Engineering, Vol. 15(1), pp. 240-243.

Ballou, D., Wang, R., Pazer, H. & Tayi, G.K. (1998). Modeling Information Manufacturing Systems to Determine Information Product Quality, Management Science, Vol. 44(4), pp. 462-484.

Batini, C., Cappiello, C., Francalanci, C. & Maurino, A. (2009). Methodologies for data quality assessment and improvement, ACM Computing Surveys (CSUR), Vol. 41(3), pp. 1-52.

Blake, R. & Mangiameli, P. (2011). The Effects and Interactions of Data Quality and Problem Complexity on Classification, Journal of Data and Information Quality (JDIQ), Vol. 2(2), pp. 1-28.

Bors, C., Gschwandtner, T., Kriglstein, S., Miksch, S. & Pohl, M. (2018). Visual Interactive Creation, Customization, and Analysis of Data Quality Metrics, Journal of Data and Information Quality, Vol. 10(1), pp. 1-26.

Boyadzhieva, D. & Kolev, B. (2010). Intuitionistic Fuzzy Data Quality Attribute Model and Aggregation of Data Quality Measurements, in: Anonymous (ed.), Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 383-395.

Bronselaer, A., Nielandt, J. & De Tré, G. (2018b). An incremental approach for data quality measurement with insufficient information, International Journal of Approximate Reasoning, Vol. 96, pp. 95-111.

Bronselaer, A., De Mol, R. & De Tre, G. (2018a). A Measure-Theoretic Foundation for Data Quality, IEEE Transactions on Fuzzy Systems, Vol. 26(2), pp. 627-639.

Caballero, I., Verbo, E., Calero, C. & Piattini, M. (2007). A Data Quality Measurement Information Model Based On ISO/IEC 15939, Cambridge, MA, pp. 393-408.

Cappiello, C., Francalanci, C. & Pernici, B. (2004). Data quality assessment from the user's perspective, Proceedings of the 2004 international workshop on information quality in information systems, ACM, pp. 68-73.

Dai, W., Wardlaw, I., Cui, Y., Mehdi, K., Li, Y. & Long, J. (2016). Data profiling technology of data governance regarding big data: Review and rethinking, Advances in Intelligent Systems and Computing, pp. 439-450.

Dorr, B. & Murnane, R. (2011). Using Data Profiling, Data Quality, and Data Monitoring to Improve Enterprise Information, Software Quality Professional, Vol. 13(4), pp. 9-18.

Eppler, M. & Helfert, M. (2004). A classification and analysis of data quality costs, Proceedings of the Ninth International Conference on Information Quality (ICIQ-04), pp. 311-325.

Even, A. & Shankaranarayanan, G. (2009). Dual Assessment of Data Quality in Customer Databases, Journal of Data and Information Quality (JDIQ), Vol. 1(3), pp. 1-29.

Even, A. & Shankaranarayanan, G. (2007). Utility-driven assessment of data quality, ACM SIGMIS Database: the DATABASE for Advances in Information Systems, Vol. 38(2), pp. 75-93.

Even, A. & Shankaranarayanan, G. (2005). Value-Driven Data Quality Assessment, Proceedings of the 2005 International Conference on Information Quality (MIT IQ Conference).

Fisher, C.W., Lauria, E.J.M. & Matheus, C.C. (2009). An Accuracy Metric: Percentages, Randomness, and Probabilities, Journal of Data and Information Quality (JDIQ), Vol. 1(3), pp. 1-21.

Heinrich, B., Klier, M. & Kaiser, M. (2009). A Procedure to Develop Metrics for Currency and its Application in CRM, Journal of Data and Information Quality (JDIQ), Vol. 1(1), pp. 1-28.

Heinrich, B., Hristova, D., Klier, M., Schiller, A. & Szubartowicz, M. (2018a). Requirements for Data Quality Metrics, Journal of Data and Information Quality (JDIQ), Vol. 9(2), pp. 1-32.

Heinrich, B., Kaiser, M. & Klier, M. (2007). How to measure data quality? A metric-based approach, Twenty Eighth International Conference on Information Systems, Montreal, pp. 1-15.

Heinrich, B., Klier, M., Schiller, A. & Wagner, G. (2018b). Assessing data quality – A probability-based metric for semantic consistency, Decision Support Systems, Vol. 110 pp. 95-106.

Hirsjärvi, S., Remes, P. & Sajavaara, P. (2007). Tutki ja kirjoita, 13. osin uud. laitos. ed. Tammi, Helsinki, 448 p.

Laihonen, H., Hannula, M., Helander, N., Ilvonen, I., Jussila, J., Kukko, M., Kärkkäinen, H., Lönnqvist, A., Myllärniemi, J. & Pekkola, S. (2013). Tietojohtaminen, Tietojohtamisen tutkimuskeskus NOVI, Tampereen Teknillinen Yliopisto, 84 p.

Lee, Y.W., Pipino, L., Strong, D.M. & Wang, R.Y. (2004). Process-Embedded Data Integrity, Journal of Database Management (JDM), Vol. 15(1), pp. 87-103.

Lee, Y.W., Strong, D.M., Kahn, B.K. & Wang, R.Y. (2002). AIMQ: a methodology for information quality assessment, Information & Management, Vol. 40(2), pp. 133-146.

Liu, Z., Chen, Q. & Cai, L. (2018). Application of Requirement-oriented Data Quality Evaluation Method, 2018 19th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), IEEE, pp. 407-412.

Loshin, D. (2001). Enterprise Knowledge Management: The Data Quality Approach, 1st ed. Morgan Kaufmann, United States, 493 p.

Mahanti, R. (2014). Critical Success Factors for Implementing Data Profiling: The First Step Toward Data Quality, Software Quality Professional, Vol. 16(2), pp. 13-26.

McGilvray, D. (2008). Executing Data Quality Projects, 1st ed. Morgan Kaufmann, 352 p.

Pipino, L.L., Lee, Y.W. & Wang, R.Y. (2002). Data quality assessment, Communications of the ACM, Vol. 45(4), pp. 211-218.

Redman, T.C. (1998). The impact of poor data quality on the typical enterprise, Communications of the ACM, Vol. 41(2), pp. 79-82.

Saunders, M., Lewis, P. & Thornhill, A. (2009). Research methods for business students, 5th ed. Prentice Hall, Harlow, 614 p.

Sebastian-Coleman, L. (2013). Measuring Data Quality for Ongoing Improvement, 1st ed. Morgan Kaufmann, 376 p.

Tuomi, J. & Sarajärvi, A. (2009). Laadullinen tutkimus ja sisällönanalyysi, 5., uud. laitos. ed. Tammi, Helsinki, 175 p.

Umar, A., Karabatis, G., Ness, L., Horowitz, B. & Elmagardmid, A. (1999). Enterprise Data Quality: A Pragmatic Approach, Information Systems Frontiers, Vol. 1(3), pp. 279-301.

Vilkka, H. (2015). Tutki ja kehitä, 4. uudistettu painos. ed. PS-kustannus, Jyväskylä [viitattu 30.01.2019]. Saatavissa: https://www.ellibslibrary.com/book/978-952-451-756-0

Wand, Y. & Wang, R.Y. (1996). Anchoring data quality dimensions in ontological foundations, Communications of the ACM, Vol. 39(11), pp. 86-95.

Wang, R.Y., Storey, V.C. & Firth, C.P. (1995). A framework for analysis of data quality research, IEEE Transactions on Knowledge and Data Engineering, Vol. 7(4), pp. 623-640.

Wang, R.Y. & Strong, D.M. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers, Journal of Management Information Systems, Vol. 12(4), pp. 5-33.

Wang, R.Y. (1998). A product perspective on total data quality management, Communications of the ACM, Vol. 41(2), pp. 58-65.

Watts, S., Shankaranarayanan, G. & Even, A. (2009). Data quality assessment in context: A cognitive perspective, Decision Support Systems, Vol. 48(1), pp. 202-211.

Woodall, P., Borek, A. & Parlikad, A.K. (2013). Data quality assessment: The Hybrid Approach, Information & Management, Vol. 50(7), pp. 369-382.

Yin, R.K. (2003). Case study research: design and methods, 3rd ed. Sage Publications, Thousand Oaks (CA), 179 p.

LIITE A: HAASTATTELURUNKO

1. Ajantasaisuus
1.1 Onko data riittävän ajantasaista?

2. Asiaankuuluvuus
2.1 Onko data hyödyllistä?
2.2 Soveltuuko data työhösi?

3. Esityksen johdonmukaisuus
3.1 Onko data esitetty johdonmukaisesti samassa muodossa?

4. Esityksen ytimekkyys
4.1 Onko data esitetty tiiviisti?

5. Helppokäyttöisyys
5.1 Onko dataa helppo muokata vastaamaan tarpeita?
5.2 Onko data helposti yhdistettävissä muun datan kanssa?

6. Maine
6.1 Onko datalla hyvä maine?
6.2 Tuleeko data hyvistä lähteistä?

7. Objektiivisuus
7.1 Onko data objektiivisesti kerätty?
7.2 Perustuuko data faktoihin?

8. Oikea-aikaisuus
8.1 Onko data käytettävissä oikeaan aikaan?

9. Saatavuus
9.1 Onko data helposti saatavissa?
9.2 Onko data nopeasti saatavissa tarvittaessa?

10. Sopiva määrä
10.1 Onko dataa riittävä määrä tarpeisiin?

11. Tulkittavuus/Ymmärrettävyys
11.1 Onko datan merkitys helposti tulkittavissa ja ymmärrettävissä?

12. Turvallisuus
12.1 Onko data suojattu luvattomalta pääsyltä?
12.2 Onko data saavutettavissa vain tarkoitetuille henkilöille?

13. Täydellisyys

13.1 Sisältääkö data kaikki tarvittavat arvot?
13.2 Onko data riittävän täydellistä tarpeisiin?

14. Uskottavuus
14.1 Onko data uskottavaa?

15. Virheettömyys
15.1 Onko data virheetöntä?
15.2 Onko data tarkkaa/oikeellista?

LIITE B: AIMQ-KYSELYLOMAKE

Käänteiset väitteet ovat esitetty “(K)”-merkinnällä.

Saatavuus
Tämä data on helposti haettavissa.
Tämä data on helposti saatavissa.
Tämä data on helposti saavutettavissa.
Tämä data on nopeasti saatavissa tarvittaessa.

Sopiva määrä
Dataa on riittävä määrä tarpeisiimme.
Datan määrä ei vastaa tarpeitamme. (K)
Datan määrä ei riitä tarpeisiimme. (K)
Dataa ei ole liikaa eikä liian vähän.

Uskottavuus
Tämä data on uskottavaa.
Tämä data on kyseenalaisen uskottavaa. (K)
Tämä data on luotettavaa.

Täydellisyys
Tämä data sisältää kaikki tarvittavat arvot.
Tämä data on epätäydellistä. (K)
Tämä data on täydellistä.
Tämä data on riittävän täydellistä tarpeisiimme.
Tämä data kattaa tehtäviemme tarpeet.

Esityksen ytimekkyys
Tämä data on muotoiltu kompaktisti.
Tämä data on esitetty tiiviisti.
Tämä data on esitetty kompaktissa muodossa.
Tämän datan esitys on kompakti ja tiivis.

Esityksen johdonmukaisuus
Tämä data on johdonmukaisesti esitetty samassa muodossa.
Tämä data ei ole esitetty johdonmukaisesti. (K)
Tämä data on esitetty johdonmukaisesti.

Helppokäyttöisyys
Tämä data on helposti muokattavissa vastaamaan tarpeitamme.
Tämä data on helposti yhdistettävissä.
Tämä data on vaikeasti muokattavissa vastaamaan tarpeitamme. (K)
Tämä data on vaikeasti yhdistettävissä. (K)
Tämä data on helposti yhdistettävissä muun datan kanssa.

Virheettömyys
Tämä data on virheetöntä.
Tämä data on virheellistä. (K)
Tämä data on tarkkaa.
Tämä data on luotettavaa.

Tulkittavuus
Tämän datan merkitystä on helppo tulkita.
Tätä dataa on vaikea tulkita. (K)
Tämä data on helposti tulkittavissa.
Tämän datan mittausyksiköt ovat selkeitä.

Objektiivisuus
Tämä data on objektiivisesti kerätty.
Tämä data perustuu faktoihin.
Tämä data on objektiivista.
Tämä data esittää tasapuolisen näkymän.

Asiaankuuluvuus
Tämä data on hyödyllistä työhömme.
Tämä data on merkityksellistä työhömme.
Tämä data on sopivaa työhömme.
Tämä data on soveltuvaa työhömme.

Maine
Tämän datan laadun maine on huono. (K)
Tällä datalla on hyvä maine.
Tämän datan laadun maine on hyvä.
Tämä data tulee hyvistä lähteistä.

Turvallisuus
Tämä data on suojattu luvattomalta pääsyltä.
Tätä dataa ei ole suojattu riittävällä turvallisuudella. (K)
Tämän datan pääsynhallinta on riittävän rajoitettu.
Tämä data on saavutettavissa vain tarkoitetuille henkilöille.

Oikea-aikaisuus
Tämä data on riittävän ajankohtaista työhömme.
Tämä data ei ole riittävän oikea-aikaista. (K)
Tämä data ei ole riittävän ajankohtaista työhömme. (K)
Tämä data on riittävän oikea-aikaista.
Tämä data on riittävän ajantasaista työhömme.

Ymmärrettävyys
Tämä data on helposti ymmärrettävissä.
Tämän datan merkitystä on vaikea ymmärtää. (K)

Tämän datan merkitys on helposti ymmärrettävissä.